EXTRATERRESTRIAL CIVILIZATIONS

Problems of Interstellar Communication

S. A. KAPLAN (Ed.)

S. A. Kaplan, Editor

EXTRATERRESTRIAL CIVILIZATION

Problems of Interstellar Communication

Translated from Russian

Published for the National Aeronautics and Space Administration
and the National Science Foundation, Washington, D.C.
by the Israel Program for Scientific Translations

S. A. KAPLAN, Editor

EXTRATERRESTRIAL CIVILIZATIONS

Problems of Interstellar Communication

(Vnezemnye tsivilizatsii.
Problemy mezhzvezdnoi svyazi)

Izdatel'stvo "Nauka"
Glavnaya Redaktsiya
Fiziko-Matematicheskoi Literatury
Moskva 1969

Translated from Russian

Israel Program for Scientific Translations
Jerusalem 1971

TT 70-50081
NASA TT F-631

Published Pursuant to an Agreement with
THE NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
and
THE NATIONAL SCIENCE FOUNDATION, WASHINGTON, D. C.

IPST Cat. No. 5780

Translated by IPST staff

Printed in Jerusalem by Keter Press
Binding: Wiener Bindery Ltd., Jerusalem

Available from the
U. S. DEPARTMENT OF COMMERCE
National Technical Information Service
Springfield, Va. 22151

X/16/3

Table of Contents

Introduction

EXOSOCIOLOGY — THE SEARCH FOR SIGNALS FROM EXTRATERRESTRIAL CIVILIZATIONS

The search for signals from extraterrestrial civilizations is one of the most intriguing problems raised by modern science. What are these signals, where and how are we to look for them, and should we devote time and effort to this search? These questions were originally in the domain of science fiction, and it is only recently that they began to be considered seriously by astronomers, physicists, biologists, linguists, and philosophers in scientific conferences and in various articles and books. However, despite the numerous questions raised and the various hypotheses advanced, there has been very little real scientific research in this direction. Even the cardinal question of the actual outcome of the encounter of mankind with extraterrestrial civilization — whether it will be beneficial or harmful — has not been answered unanimously. It suffices to mention how the excessively optimistic prospects of interstellar communication drawn by I. A. Efremov in his "Andromeda Nebula" contrast with the distressing picture envisaged by F. Hoyle and Ch. Elliott in their "A for Andromeda." Incidentally, both these books were written by eminent scientists, fully conversant with the grave implications of the problem, and not by professional science-fiction writers who sometimes stand accused of flippant treatment of the subject.

The expansion of man into outer space led to a rapid development of new branches of science and technology. One of these new disciplines is exobiology, a science dealing with the origin and evolution of life under extraterrestrial conditions. The very wide range of topics considered by exobiology attracted the attention of scientists from a variety of fields. Some problems of exobiology are even now nearing final solution, whereas others are still in the embryonic stages of research.

One of the fundamental problems of exobiology is purely astronomical: what is the probability of any individual star being surrounded by planets with life-sustaining conditions? In other words, the primary task is to find the probability of existence of a planet with a mass not radically different from the Earth mass, adequate axial rotation parameters, and an atmosphere, which lies in the "life-sustaining heat zone," i.e., not too far from the primary to be permanently frozen and yet not too near it for the surface to be scorched. Although there is a measure of uncertainty in the very formulation of the problem, a more or less definite solution has been obtained by now. The probability of the existence of such a life-sustaining planet is of the order of a few percent.

However, the existence of conditions which are potentially capable of sustaining life does not imply that life actually exists on the particular planet.

Unfortunately, there is a great deal of uncertainty in this purely biological aspect of the matter. Most authorities seem to be of the opinion that the probability of life inception is fairly high. There is, however, an alternative point of view, equally valid in terms of the actual proof available (or rather total lack thereof), which suggests that the probability of life inception is negligible even under ideally suitable conditions. The extreme difficulty of the problem is further aggravated by the lack of a reliable theory of the origin of life on Earth. The attempts to reproduce this process in laboratory have failed so far. The discovery of even minute traces of life on Venus or Mars or the demonstration of repetitiveness or multiplicity of the process of life inception on Earth would provide invaluable information toward the solution of the problem (together with laboratory research). Therefore we do not exaggerate if we say that the success of exobiology in the solution of its fundamental problem — elucidating the possibility of life originating under certain conditions — largely depends on the level of our space technology. The rapid advances of modern astronautics instill us with hope that the probability of life on a planet endowed with appropriate conditions will be determined in the nearest future.

No less complex and equally far from solution is the problem dealing with the probability of evolution of life from the inception of the most primitive life forms to intelligent beings. The various opinions here again cover a wide spectrum, ranging from the extreme suggestion that the development of intelligence is a single-valued consequence of the inception of life to less categorical statements which regard the biological evolution as a succession of critical, non-repeatable and unpredictable steps, a chain that can be severed by the slightest of chances. If we adopt the latter point of view, the life on Earth is a unique phenomenon, possible within the limits of the entire Metagalaxy. This is clearly not a very appealing assumption.

The present author, because of total lack of background in biology, will have to confine himself to an expression of hope that the fundamental problem of exobiology will find its solution in the not too distant future and that the probability of evolution of intelligence from primitive life forms is not too low.

Finally we come to the remarkable problem of the evolution of intelligent societies outside the Earth, the problem of extraterrestrial civilizations. As the emphasis here is on the evolution of society and we can essentially regard the topic as falling within the framework of a new scientific discipline, concerned with the study of hitherto undiscovered societies, we would use the term "exosociology" for this discipline, by analogy with exobiology. I. S. Shklovskii's suggestion, "cosmosophy" /1/, is somewhat inconvenient in our opinion and does not fully reflect the true task before us.

No science can be nourished by purely theoretical, "cosmosophic" concepts, and exosociology is no exception to this rule. Experiments and observations are essential components of any science. At the present stage of its development, exosociology can draw for experimental data upon the only civilization known to us, the Earth civilization. The real and significant observational fabric of exosociology will be provided by the analysis of signals from extraterrestrial civilizations, assuming that such signals will be detected. It is this basic assumption that is reflected in the title of the Introduction.

Exosociology is the subject of the present book. Exobiological topics, i.e., problems relating to the origin and the evolution of life in outer space,

are not considered. A fairly extensive literature is available at present (see, e.g., /1/ and /2/).

The reader may naturally question the need and the urgency of a special volume on exosociology at the present stage, when no systematic search for signals from extraterrestrial civilizations has begun and the chances of discovery of these signals are not very high. It is our belief, however, that a book of this kind is urgently needed, and this for the following reasons. First, systematic search for signals from extraterrestrial civilizations will eventually be organized, and it is better to be prepared with all the necessary theoretical and practical background information relating to this search. Second, exosociological research may yield certain "byproducts" which will be of considerable significance for "terrestrial" science. For example, the search for radio sources of suspected artificial origin is entirely analogous to certain problems of modern radio astronomy, and at first glance has no relation to exosociology. The decoding of messages from outer space may provide much valuable information relating to pure linguistic problems. And so far we did not mention the forecasting of the future growth and development of civilizations. Therefore, having organized a systematic search for signals from extraterrestrial civilizations and proceeding with a research into the various problems of exosociology, we will not end up losers even if no extraterrestrial signals are detected in the near future. The potential gain, on the other hand, is hardly imaginable.

The six chapters of the book deal with various aspects of the search for signals from extraterrestrial civilizations. To help the reader, we will try to present a general survey of the problem and the current view of its basic aspects.

The theory of development of civilizations

The Earth civilization — the only known example of a society of intelligent beings — has existed for a very brief period of time on the astronomical time scale, for no more than a few millennia. The time interval accessible to actual research is even smaller. And yet, the main topic of exosociology is the study of civilizations over the entire span of their evolution, which, at least in principle, may be comparable with the astronomical time scale (millions and billions of years). In any case, signals can be detected only from civilizations markedly exceeding the level of development of the Earth civilization.

Exocosiology should thus be able to study supercivilizations, i.e., the evolution of intelligent societies over very long, astronomical periods of time.

It would seem that the solution of this problem should start with a detailed forecast of the further growth of the Earth civilization. However, this immediately leads us to a fundamental difficulty. Any forecast is essentially based on an extrapolation of previous development. This extrapolation is evidently valid over a period which is at most comparable to, and usually much smaller than, the period of time on which the forecast is based. It is not by chance that most forecasts of the future of mankind are limited to the year 2000 (occasionally venturing to the year 2100)!

The intrinsic imperfection of the extrapolatory approach emerges from the fact that its automatic application to the forecasting of the future development of the Earth civilization inevitably leads to so-called "explosions" — very rapid growth of some indices.

Probably the best known example is the "demographic explosion" or the "population explosion," i.e., the conclusion that the Earth population will become infinitely large around the years 2020—2030. Another example is the "energy" or "power" explosion. Calculations show that around the year 2100, the power production on the Earth will reach such a level that the temperature of the planet will increase indefinitely. Finally, we seem to be on the threshold of the so-called "information explosion," when the volume of information accumulated by science will become infinite (this event is "scheduled" for around the year 1980).

There is no doubt that none of these explosions will actually occur, but it is not clear how the "critical" moments will be avoided and how the growth characteristics will change to prevent the crisis. Repopulation of mankind in outer space is often proposed as a universal remedy. A simpler solution will probably present itself when the time is ripe. Analysis of the succession of the growth characteristics is thus one of the principal problems to be tackled in forecasting the future development of civilization (see Chapter V).

Thus, despite the considerable interest attached to the forecasts of growth of the Earth civilization, their contribution to exosociology is negligible. For this reason, we will not go into these forecasts in any detail, and we would only like to mention that according to A. Clarke /6/ and the forecasts developed by the Rand Corporation in the USA, the encounter with extraterrestrial civilizations is deferred to the second half of the 21st century.

It therefore seems that at this stage it is more advisable to start looking for general laws governing the development of intelligent societies and civilizations in some more abstract form, based on the modern cybernetic concepts of complex systems. We should try to evolve general definitions of the concept of civilization and to analyze the evolutionary trends emerging from this system-theoretical definition. The following definition of a civilization is advanced in Chapter I: "A highly stable state of matter capable of acquisition, abstract analysis, and application of information for the purpose of extracting the maximum quantity of information about the environment and itself and developing survival reactions." Chapter V mentions another general feature: "Simple systems evaluate these outside stimuli only in order to determine the state of the internal and the external media at the material time, whereas more complex systems can respond to a forecast future state of the environment as predicted on the basis of the current measurements." Proceeding from these definitions, we can expect an unlimited development of civilizations and an intrinsic tendency to establish contact with one another. The cybernetic approach to the problem of supercivilizations is discussed in more detail in Chapter VI.

We are not only very far from the solution of the fundamental problem of exosociology, i.e., the elucidation of the general laws governing the development of civilizations as intelligent societies, but we still have not formulated this problem in precise terms. It is our belief, however, that the considerations presented in Chapters I, VI, and partly V will help in this direction.

Note that the establishment of contact with extraterrestrial civilizations may not only lead to radical changes in our basic concepts regarding the intelligent society, however "logical" these concepts had appeared prior to the encounter with the other civilization, but also greatly affect the future development of our own civilization. This will be the result of the "feedback effect," often discussed, in particular, in connection with the beneficial or harmful results of "interplanetary" encounters.

The search for signals from extraterrestrial civilizations

Despite the tremendous volume of information accumulated by modern astrophysics and radio astronomy, no such signals have been detected so far. If we remember that most discoveries are quite accidental and happen generally whenever they are least expected, there is no reason for over-optimism in this respect. It is hard to say what the exact reasons are. It may be that no other civilizations exist sufficiently close to the Sun which are capable of sending signals into outer space. And yet, most authorities are of the opinion that supercivilizations are quite abundant. We will be able to reach sound conclusions, however, only after going through a complete program of search for signals from other civilizations. This is one of the reasons for our conviction that such a search program must be launched immediately.

Incidentally, even if extraterrestrial civilizations do not send special signals into space, there is a possibility that we will be able to "intercept" their internal transmissions (television broadcasts, for instance). The artificial radio emission of the Earth has reached by now a fairly high level of intensity /1, 5/, and that of supercivilizations will be many times higher. Combination of high-sensitivity receivers with large-base interferometers (see below) will probably facilitate the problem of "interception" of the transmissions of extraterrestrial civilizations.

The program of search for signals from extraterrestrial civilizations is discussed in detail in Chapters I and III. The first step is apparently a radio survey of the sky with the aim of detecting radio sources of minimum angular dimensions. Indeed, the antennas of the sending supercivilizations, irrespective of the particular information that they transmit, will be very small compared to the astronomical scale of distances. In principle, transmitting systems of planetary size are possible, but even the planetary scale is vanishingly small compared to the size of other radio sources in space. The current resolving power of radio observations has reached 0".005. This resolution was attained with a radio interferometer using separate recording in each arm. In principle, radio-interferometric observations can now be made with a base of the order of the Earth's diameter, and in future the base will probably be increased to about 1 a.u. (giving resolution of $2 \cdot 10^{-9}$ angular seconds!).

There is a whole range of other criteria which identify the probable artificial origin of a radio source. These criteria are described and discussed in detail in Chapters I and III, and a more general aspect of the identification of artificial signals is given in Chapter VI. Regular variations in the signal, definite polarization, and other features of this kind must be

analyzed in great detail. The artificial nature of the signal can also be inferred from the statistical properties of the electrical field of the radio wave. The most reliable criterion, however, is nevertheless the exceedingly small angular size of the source.

The choice of wavelengths at which artificial sources are to be sought presents another important problem. It is generally agreed that the idea of communication with extraterrestrial civilizations passed from the domain of science fiction to the domain of science in 1959, when Cocconi and Morrison suggested that the signals of extraterrestrial civilizations should be sought at the natural wavelength standard, the 21 cm radio line of the hyperfine structure of atomic hydrogen. This suggestion naturally met with certain opposition; in particular, it has been pointed out that the interstellar medium is highly absorbing at this wavelength, so that the higher harmonics of the 21 cm line should probably be used.

There are, however, other natural wavelength standards, e.g., the radio lines of the so-called Λ-doubling of the hydroxyl molecules OH. In fact, four lines are observed, associated with the combination of Λ-doubling and the hyperfine structure. The mean wavelength of the four lines is $\lambda = 18$ cm. For all the four hydroxyl lines, the interstellar absorption is significantly lower than for hydrogen lines, but it is nevertheless quite high.

The hydroxyl radio lines have recently attracted considerable attention on the part of radio astronomers and astrophysicists, following the discovery of a "natural maser effect" at these wavelengths: very narrow (with a Doppler width corresponding to a temperature profile of a few degrees Kelvin) and very strong (with a brightness temperature of over 10^{13} deg) highly polarized hydroxyl lines have been observed for a number of sources located near the regions of hot ionized interstellar hydrogen. The unusual behavior of these lines explains their new name, the "mysterium lines." If we further remember that the radio sources of "mysterium" lines are characterized by the smallest known angular dimensions, of the order of a few thousandths of an angular second (this corresponds to linear dimensions of a few astronomical units for their distances from the Earth), no wonder that these sources are suspected as being of artificial origin.

We are far from suggesting that the "mysterium" sources are extraterrestrial civilizations, but this example clearly illustrates the great importance of detailed observations and analysis of all the "suspicious" objects.

Further note that at centimeter and decimeter wavelengths, which are the most suitable for purposes of interstellar radio communication (the interstellar noise is the least at these wavelengths, see Chapters I and II), there are other molecular lines which in principle can be used for signal transmission by extraterrestrial civilizations. Finally, radio transmission is also possible and even highly probable in the continuous spectrum between 10 and 50 cm wavelengths, and this wide frequency band ensures a sufficiently high rate of information transmission (Chapters I and III).

Recently considerable attention has been attracted by the discovery, on 6 August 1967, of the so-called "pulsars," pulsating radio sources with a remarkably regular periodicity of pulse repetition in a continuous spectrum.

The observations of pulsars in the first months following their discovery was closely linked with the problem of search for signals from extraterrestrial civilizations. We will therefore consider this chapter of science in

some detail. The name pulsars was assigned to certain objects which emit discrete and very short pulses (with a duration of the order of a few hundredths and even thousandths of a second) in a wide region of the continuous radio spectrum. In the intervals between the successive pulses, no pulsar emission has been observed so far. The radio pulses differ in shape and in amplitude, i.e., in emitted radio power. The pulses reveal a certain fine structure: those of numerous pulsars are made up of so-called subpulses. The pulses of different pulsars have different shapes, and even the pulses of one pulsar are variable in this respect. The magnitude of pulsars is variable between even wider limits, and occasionally they vanish altogether. In many, though not in all, cases, the pulsar pulses are polarized. At least some of the pulsars probably emit pulsed radiation in the visible spectrum also. The various features described so far are quite usual for natural astrophysical sources, and possibly even for ordinary stars. Certain features of the pulsar radio emission are quite similar to the sporadic radio emission of the Sun. However, one of the pulsar properties — in fact, their main property which is responsible for their very name — appeared highly unusual. The pulses revealed a strikingly regular periodicity of recurrence. The first of the discovered pulsars showed pulse recurrence periods close to 1 sec, and the exact period of each pulsar remained constant with astonishing precision: over a year, the period did not change to the seventh or eighth position after the decimal point. For example, the period of the best known pulsar, CP 19019, is 1.33730109 ± 10^{-8} sec. Soon after that, it was established that the pulsar periods systematically increase (the change is in the seventh significant digit during one year). This strict periodicity led A. Hewish, who headed the group responsible for the discovery of pulsars in Cambridge (England), to the suggestion of the possible artificial origin of pulsars. The press at that time succinctly described the pulsars as the signals of the "little green men." A. Hewish kept the discovery as a closely guarded secret for about six months after the observation of the first pulsar, a highly unusual development in the modern scientific community. It was only after the discovery of three other pulsars in Cambridge that the results were announced. Almost simultaneous discovery of several extraterrestrial civilizations is a highly unlikely event.

Note that the existence of a strict periodicity in natural processes which take place in astronomical objects is by no means an unusual phenomenon. Obvious examples are the axial rotation periods of planets or binaries. Certain variable stars (the relatively small group of RR Lyrae stars, typical type I population stars) are distinguished by exceptional stability of light variation: their periods do not change significantly over a million cycles. So far, however, the astronomers have dealt with periods measured in hours and days, whereas in pulsars the characteristic periods are seconds or fractions of a second, but this does not appear to be a fundamental distinction.

Besides strict periodicity, the pulsars show nothing that supports the hypothesis of artificial origin (see Chapters I and III). This hypothesis survived for a few months only. By the end of 1968, 27 pulsars had been discovered with periods ranging from 300 to 3 seconds. The properties of pulsars proved to be highly interesting and highly unusual: some theories identify these objects with spinning neutron stars (these theories explain both the strict periodicity and the increase in period). However, the pulsars can be said to definitely fall outside the scope of our book.

The modern theory of communication enables us to analyze the conditions of signal transmission through interstellar space, to consider the requirements to be met by the transmitting and, especially, the receiving systems and antennas. This analysis, carried out in considerable detail in Chapter III, will help to select the optimum antenna parameters, receiver band widths, and scanning periods in connection with the program of search for extraterrestrial signals. We would only like to stress that the main problem falls into two separate parts: the direct search for signals ("discovery of artificial sources") and reception of information from extraterrestrial civilizations. For straightforward detection purposes, the useful signal may be much weaker than the noise level. These signals can be picked up with the aid of averaging techniques (as is often done in radio astronomy), but part of the information is naturally lost in the process. If we are interested in merely detecting signals from extraterrestrial civilizations, without interpreting their meaning, the "power" of the civilizations may be several orders of magnitude less than in cases when full reception of information is required (and the maximum distances are correspondingly larger). This means, incidentally, that the first instances of signal detection from extraterrestrial civilizations will not lead to catastrophic consequences.

We do not intend to present here any specific programs of search for extraterrestrial civilizations. The actual program will be decided upon only after a comprehensive and all-sided analysis of the possibilities of modern radio-astronomical equipment, taking into consideration the actual observation time available on the largest radio telescopes for this project. The use of radio interferometers with a base comparable to the Earth's diameter will be impossible without close international cooperation on the project.

The authors nevertheless hope that the analysis of the problem of search for signals from extraterrestrial civilizations, presented in this book, will promote the development of a large and comprehensive program with higher chances of success than the well-known Ozma project initiated by F. Drake in 1956 for detailed observations of the two close neighbors of the Sun, ε Eridani and τ Ceti.

Decoding aspects of the program of search for extraterrestrial civilizations

Before any signals have been received, we are in no position to discuss their probable information content. There is absolutely no point in trying to guess now whether these will be television images (the most comprehensible language, at least from our point of view) or messages based on the principles of formal logic, akin to the famous LINCOS language, or perhaps something entirely different.

It nevertheless seems that we are ripe for a precise formulation of certain basic problems relating to the decoding of unknown messages. Consider one example. Suppose a certain message has been received; let this be a text written in an unknown language, with an unknown alphabet and unknown rules for division into sentences and words; even the letter codes are unknown. The only available piece of information is that we have received a sequence of signals, e.g., pulses, of definitely artificial origin. Can this text be decoded so as to disclose its meaning and contents? For

purposes of decoding, it is necessary (though not sufficient) to determine the letter codes and the division into words and sentences, to establish the grammar of the language, to compile a dictionary, and to elucidate the pronunciation of the letters and the words.

Consider another example. A fragmentary message (e.g., distorted by noise) has been received, but it is almost certainly a part of an image (a static television picture). Can we reconstruct the entire picture from the received message, i.e., determine the number of lines and scanning elements in each line? The best-known example of messages of this kind is Drake's cosmogram (described in Chapter IV), in which a sequence of 1271 elements (ones and zeros) is used to code the picture of certain creatures (remarkably like human beings, only somewhat taller) inhabiting the fourth planet of some planetary system. The deciphering of this cosmogram is greatly facilitated by the fact that the number 1271 can be split either into 31 lines of 41 elements each, or into 41 lines of 31 elements each. There are thus two alternative solutions, and the right answer is almost obvious. However, if we miss a few of the first elements of the message, the screen is no longer rectangular and the message will probably be undecipherable.

There is, of course, a possibility that the signals from extraterrestrial civilizations contain the key for the decoding of the transmitted message. The question is directly related to the topic of call signals, which should identify the artificial origin of the signals. This idea opens wide horizons for various assumptions and speculations. We will consider the problem of call signals and simple keys for decoding in Chapters I, III, IV, and VI. In our opinion, however, it is better and more worthwhile to concentrate on the problem of decoding of unknown messages assuming total absence of any decoding keys. This constitutes the topic of Chapter IV, which was written by a professional linguist.

The method of decoding proposed in this book essentially amounts to what is known in physics as the method of construction of correlation functions (they are called quality functions in Chapter IV) for messages. Indeed, certain combination rules exist for the consonants and the vowels, for words which belong to different grammatical classes, and correlation functions constructed for different symbols of the received message therefore provides certain identifying information about these symbols. If the message comprises the scanning elements of a picture, the correlation function permits reconstructing the successive lines and then the entire picture. This decoding procedure naturally involves a large volume of computations, and therefore it must be handled by computers. The problem of decoding thus reduces to a construction of an algorithm for the computation of correlation functions and their comparison with certain criteria (of the type of the entropy criterion) which make it possible to select the best solution (the entropy of ordered distributions is minimum). It is moreover clear that since the decoding procedure is based on statistical processing, a sufficiently large sample, i.e., a sufficiently long message, is needed for the decoding to prove effective in complex cases. Simple examples nevertheless can be solved using short messages.

We would like to stress that Chapter IV mainly deals with the decoding of messages from the linguist's point of view. The reader interested in the general principles of decoding may read only the first seven sections. The remaining four sections contain various algorithms intended for the solution

of more complex problems. Despite the sophisticated algorithms, however, we are still very far from complete decoding of long texts in an unknown language. Yet the principles have crystallized, and the rest is a technical matter.

* * *

We tried to present a brief survey of a new scientific discipline — exosociology, the search for signals from extraterrestrial civilizations — and at the same time review the contents of this book. I would now like to add a few comments in my capacity as the editor of this volume.

The original intention was that each chapter should embrace one well-defined aspect of the problem of search for signals from extraterrestrial civilizations. The result is thus not a collection of papers, but a kind of monograph. The main difficulty, however, is that exosociology, like any new scientific discipline, still gropes uncertainly among differences of opinion and lack of firmly established concepts. Even the different contributors to this volume differ in their opinion on certain subjects. It was not the editor's intention to impose his own point of view upon the authors or to act as an arbitrator. As a result, however, a number of topics, e.g., the concept of a civilization, the date of the energy explosion, etc., are discussed in different chapters, sometimes from different points of view. The reader will have to decide for himself whose arguments sound the most convincing. He may even feel free to form his own opinion on the subject.

It should be emphasized, however, that these "differences of opinion" are relatively few and, on the whole, the contributors have pursued the original aim, namely a scientific discussion of the problem of search for signals from extraterrestrial civilizations on the modern level, in order to stimulate further interest in this problem.

The book is intended for a wide audience, although it is not a popular book in the usual sense of this word The authors did their best to maintain a high scientific level in their presentation, without going into tedious technical details which are of interest to narrow specialists only (the only exception to this rule is probably the second part of Chapter IV). The main difficulty for the reader is the great variety of subjects covered: radio astronomy, theory of information, linguistics, cybernetics, aspects of civilization . . . .

Some readers will probably feel that certain sections are much too superficial, whereas others are excessively detailed. Certain chapters are too simplified, and others are too complicated. In partial justification of this, we would like to point out that it is very difficult to maintain a consistently uniform level of presentation in a volume written by a team of contributors on such a wide spectrum of subjects.

The present book is radically different from previous publications on the subject of extraterrestrial civilizations. References /3/ and /4/, for example, are collections of articles and papers, and therefore do not provide a comprehensive picture of the problem. Moreover, they are largely outdated by now.

W. Sullivan's book is more of a popular discussion of the various events associated with the problems of exobiology, and thus does not provide a consistent analysis of the fundamental problems.

I. S. Shklovskii's book /1/ is unquestionably of the greatest interest. Unfortunately, it was written quite a number of years ago and numerous aspects of the problem of extraterrestrial civilizations are therefore not mentioned. Furthermore, the presentation is much more popularized than in the present volume.

It would seem that the present volume is the first scientific monograph in the literature on the subject of search for signals from extraterrestrial civilizations.

In conclusion, all the contributors would like to acknowledge the great help of Acad. V. A. Kotel'nikov for valuable suggestions that helped to improve the finished product, and especially the assistance of L. I. Gudzenko, who read through the entire manuscript and offered numerous comments concerning the general presentation and the particular problems discussed.

Bibliography

1. Shklovskii, I. S. Vselennaya, zhizn', razum (Life and Intelligence in the Universe). 2nd Ed. – "Nauka." 1965.
2. Sullivan, W. We are not Alone. – McGraw-Hill. 1966.
3. Cameron, A. (Editor). Interstellar Communication. – New York. Benjamin. 1963.
4. Vnezemnye tsivilizatsii (Extraterrestrial Civilizations). Proceedings of a Conference.* Byurakan, 20–23 May 1964. – Izd. AN Arm. SSR. 1965.
5. Kaplan, S. A. Elementarnaya radioastronomiya (Elements of Radio Astronomy). – "Nauka." 1966.
6. Clarke, A. C. Profiles of the Future. – Harper and Row. 1962.

* [English translation published by Israel Program for Scientific Translations, Jerusalem, IPST Cat.No.1823, NASA TT F-438 TT 67-51373.]

Chapter I

THE ASTROPHYSICAL ASPECT OF THE SEARCH FOR SIGNALS FROM EXTRATERRESTRIAL CIVILIZATIONS

§1. INTRODUCTION

The search for extraterrestrial civilizations is intimately linked with the principal problems of modern astrophysics. Let us try to establish what part of the proposed search program actually coincides with astrophysical research and what the specific requirements of the observations in this program are.

Accurate long-range prediction of the principal problems and the directions of development of space science is a fairly difficult problem. The current tendencies, however, which will leave their indelible imprint on the next few years are quite obvious.

In the next 5 — 10 years, all the radiation sources with the largest observable flux in every region of the electromagnetic spectrum will have been discovered and studied to a certain extent (**A**).* This is a realistic goal in view of the development of electromagnetic radiation detectors, i.e., radio receivers, bolometers, photosensitive detectors and materials, and photon counters. The sensitivity of these devices will soon reach the natural limit (in some spectral regions, this limit sensitivity has been attained already, e.g., the modern photon counters used in measurements of X-ray radiation from outer space detect every single impinging quantum). When the limit sensitivity is attained, we will be able to cover various cosmic objects in the entire electromagnetic spectrum, and thus virtually all the astrophysical information contained in cosmic radiation. We are thus nearing the solution of a highly important astrophysical problem:

Identification and exploration of the main (in terms of some parameter) cosmic objects (primarily objects of maximum luminosity, or radiation power, in a given spectral range, objects of the largest mass, and objects which account for the bulk of matter in the Universe) (**B**).

The primary problem of this exploratory trend is the determination of the luminosity function $N_L(L_\nu)$ and the mass function $N_M(M)$ of all the objects, where L_ν is the spectral power radiated by the object. Unfortunately, the solution of problem **A** does not imply a simultaneous solution of problem **B**

* The main propositions of this chapter are identified by bold-face letters.

(although the inverse is probably true).* Indeed, objects of the highest luminosity (e.g., supernovae, quasars) are exceptionally rare in the Universe. Therefore, the mean distance between these objects (and hence the most probable distance to the nearest source) is tremendous, and these high-luminosity objects may not be the brightest. The nearest quasar 3C 273 has a brightness of 12.5 stellar magnitudes in the optical spectrum. There are over four million stars of this magnitude in the sky, and the quasar therefore escaped optical detection for a long time; in the radio spectrum, on the other hand, this quasar is one of the hundred brightest objects, and the radio astronomers noticed it immediately.

Let us estimate the observability in a given spectral range using the luminosity function. Let

$$N_L \propto L_\nu^{-n},$$

where the index n can be determined from observations. The number of sources in unit volume with luminosities between $\frac{1}{2}L_\nu$ and $\frac{3}{2}L_\nu$ is then given by

$$N_L \propto L_\nu^{1-n}.$$

The observed flux from the nearest source whose luminosity falls between $\frac{1}{2}L_\nu$ and $\frac{3}{2}L_\nu$ is

$$F_\nu \propto \frac{L_\nu}{R^2},$$

where the mean distance between the source is $R \propto N_L^{-1/3}$.

Hence,

$$F_\nu \propto L_\nu^{\frac{5-2n}{3}}. \qquad (1.1)$$

We see from this relation that if $n > 5/2$, then $\frac{5-2n}{3} < 0$, and the lowest luminosity sources prevail among the sources with the maximum observed flux; if, on the other hand, $n < 5/2$, the situation is reversed, and the maximum luminosity sources prevail among the brightest objects.

The first of the two possibilities obtains in the comparison of the mean radiation from normal galaxies, radio galaxies, and quasars. These objects are not numerous, so that n is high, and therefore the normal galaxies prove to be the brightest among all the extragalactic optical sources.

On the other hand, if we consider the optical radiation of normal galaxies only, we have $n < 5/2$ and therefore the brightest observed objects are the most powerful. A similar situation is observed for extragalactic radio sources. Figure 1 plots the radio luminosity function /46/, which shows that in a wide range of luminosities, $n \sim 2.2$, so that the brightest objects are also the most powerful, and it is these powerful sources that are mainly studied today. Unfortunately, no such analysis can be undertaken for the mass function, since no reliable data are available at this stage. The only established fact is the mass

* N_L and N_M is the number of objects in unit volume with radiation power in the range $L \pm \delta$ and mass in the range $M \pm \delta M$, respectively, where $2\delta L_\nu$ and $2\delta M$ are unit intervals of luminosity and mass.

distribution of the stars /47/. This distribution is also adequately fitted with a power function with $n \approx 2.35$.

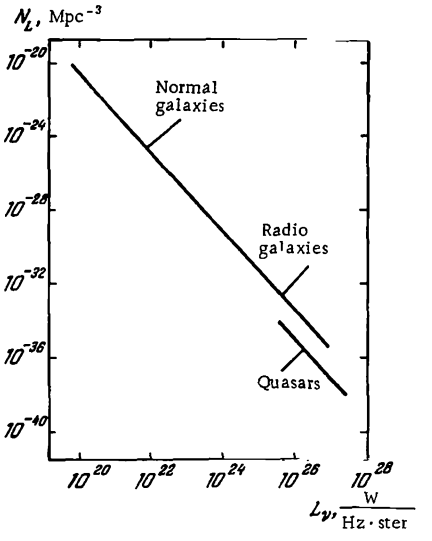


FIGURE 1. The luminosity function of extragalactic radio sources.

Problem B stresses the main tendency of development of the astrophysical research in the near future. In particular, if the activity of extraterrestrial civilizations is responsible for the radiation power of some astronomical objects, these civilizations stand a good chance of being discovered. Since in the nearest future all the regions of the electromagnetic spectrum will be accessible to space exploration, we have to prepare a suitable research program and to assess the chances of success in our search for extraterrestrial civilizations.

§2. THE MAIN DILEMMA

The main starting point for our problem probably stems from the following dilemma:

There is a high probability that civilization is a universal phenomenon, and yet there are no currently observed signs of cosmic activity of intelligent creatures (C).

Indeed, the data available on the number of planetary systems and the conditions for the evolution of life on planets suggest that life is probably a fairly commonplace and regular occurrence in the Universe. A detailed analysis of these topics will be found in /1, 2, 3/.*

* In particular, recent paleontological data convincingly prove that the inception of life on the Earth some 3 billion years ago took place simultaneously in numerous independent channels /4/.

According to most estimates, the age of our planetary system and the age of the Sun (reckoned from the time of their condensation) is from 4 to 6 billion years. It is significant that both the Sun and the planetary system are second-generation objects, but since the age of the oldest objects in the observable part of the expanding Universe (or, more precisely, the age of the first-generation objects) is at least 10 billion years, there are probably planetary systems billions of years older than the solar system. This conclusion suggests the possible existence of civilizations which are billions of years more advanced than our civilization. Taking into account the present rate of progress of our civilization, we can probably expect something nearing intentional and controlled reorganization of all matter in our part of the Universe from civilizations developing over these cosmogonic periods.

And yet, our astronomical data at first glance do not provide any indications of such cosmic activity. In our opinion, a detailed analysis of proposition C may provide the best foundation for the discussion of the program of search for extraterrestrial civilizations (EC). We will try to evaluate the various aspects of this dilemma in order to critically assess its relevance.

There may be two alternative answers resolving the dilemma:

1) either the current data on the absence of "supercivilizations" are wrong;

2) or there exists some fundamental factor slowing down the development of each and every civilization.

§3. THE COMPLETENESS AND RELIABILITY OF MODERN ASTROPHYSICAL DATA

As we have already noted, there can be no serious doubt regarding the existence of numerous planetary systems (although planets with masses of the order of the Earth's mass cannot be directly observed with modern telescopes (see /1, 2, 3/). Estimates of the number of planets which may be suitable for the evolution of life do not give any indication of the Earth's unique position in the Universe, either (see /1, 2, 3/).

The Sun and the solar system are thought to be second-generation objects, but if it were not so, there would be a definite probability of the Earth being the oldest object of this kind in the observable part of the Universe and our civilization being also the oldest.

At this point, we will have to review the current evidence relating to the age of the solar system.

Most stars whose physical parameters are close to those of the Sun remain in a steady-state condition for a long time, retaining constant radius and luminosity. The loss of radiant energy is made up by the energy released in nuclear reactions in the stellar interior. These concepts were used to develop the theory of stellar evolution according to which the steady-state phase of the Sun's evolution may take about 13 billion years, i.e., the entire evolutionary phase of the Metagalaxy. On the other hand, the age of terrestrial rocks and meteorites determined by chemical analysis of radioactive isotopes and decay products

is 4—5 billion years. This figure is usually adopted as the age of our planetary system and the Sun, since the modern theory of formation of planetary systems points to simultaneous condensation of the planets and the primary star from interstellar gas-dust clouds.

Recent results, however, seem to have substantially revised upward the age of the Earth and meteorites (see /5/). Thus, Fisher /5/ reported the results of K—Ar dating which gave an age of up to 10 billion years for some iron meteorites. The same technique gave an age of up to 10.8 billion years for terrestrial rocks /6/. Although these and other similar data by no means provide a conclusive proof of a new longer evolutionary scale of our planetary system, we cannot just ignore them.

Another aspect of this problem is related to the chemical composition of the planets. The condensation of Earth-type planets requires a sufficient content of the heavy elements in the interstellar medium, and we are thus faced with the unanswered question of the evolution of the interstellar medium and the genesis of the heavy elements in general.

In accordance with modern data on the evolution of the observable part of the Universe, it seems that all the chemical elements were formed in nuclear reactions from an original pure hydrogen plasma. Until recently, these processes were assumed to take place in stellar interiors only, the heavy elements being produced by reactions during supernova explosions. Subsequently, the heavy elements are ejected into the interstellar medium /1/. This mechanism obviously supports the hypothesis which treats the Earth-like planets as second-generation objects.

Lately, however, a new class of first-generation objects were discovered, which also show a high content of heavy elements. We mean here the quasars. The objects are primarily remarkable in that their radiation power is the highest among all the known sources of radiation in the Universe. As a result, they can be observed over tremendous distances and, because of the finite velocity of light, they provide a tool for probing into the distant past of the Universe. Figure 2 is a photograph of one of the farthest quasars 3C 9. The spectral lines of these objects show a strong red shift because of the observed expansion of the Universe. For 3C 9 the red shift is $z = \frac{\Delta\lambda}{\lambda} \approx 2$, so that all the wavelengths increase relative to the laboratory standards by a factor of $1 + z = \frac{\lambda}{\lambda_0} \approx 3$. The time between the emission and the observation of radiation for distant objects essentially depends on the particular cosmological object used. In the Einstein—de Sitter model (space curvature $k = 0$, acceleration parameter $q_0 = {}^1/_2$), the propagation time of a light signal is

$$\tau = \frac{2}{3H_0} \frac{(1+z)^{3/2} - 1}{(1+z)^{3/2}}. \qquad (1.2)$$

Here H_0 is the Hubble constant (for small red shifts z, the distance to the object is $\frac{cz}{H_0}$). The value of this constant is $H_0 \sim 30$ km/sec $\cdot 10^9$ light years. In this model, the light from 3C 9 takes about $\tau = 5.3$ billion years to reach the Earth. (The relevant data for the calculations using other models will be found in /7/.)

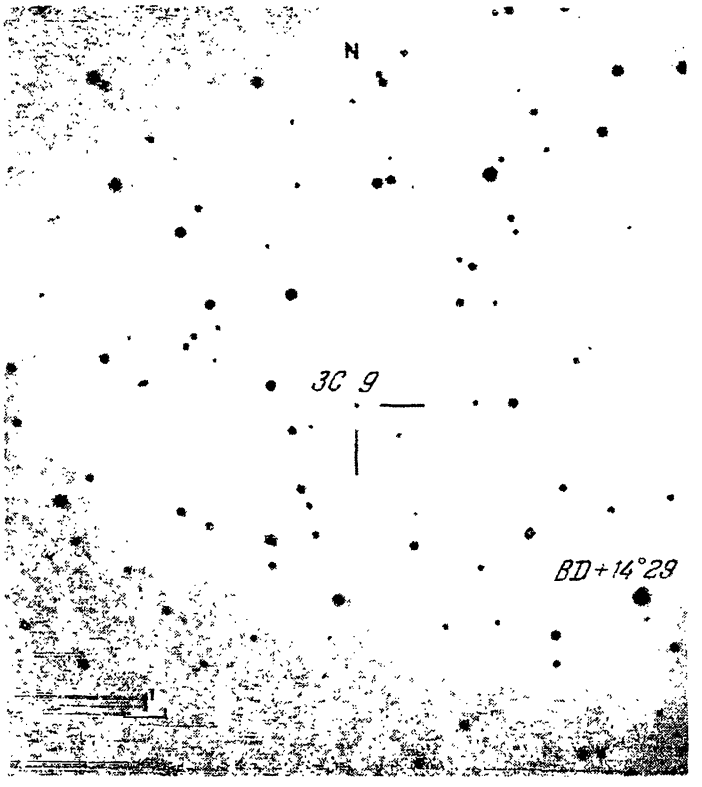


FIGURE 2. The quasar 3C 9.

The crucial point of the entire problem, as we have noted before, is the discovery of normal chemical composition in these objects /8/. In other words, the mean abundance of the chemical elements (at least of the most abundant species) in quasars is close to that observed in the neighborhood of the Sun. At the same time it has been established that quasars lie in regions where the concentration of ordinary galaxies is much below the average (between clusters of galaxies). They apparently form directly from the intergalactic medium. The heavy chemical elements are possibly synthesized in the quasar interiors, since the conditons prevailing in quasar explosions are probably even more favorable for nucleogenesis than supernova explosions. However, the similarity in the chemical composition of various quasars is really striking. It is therefore not improbable that the heavy elements were synthesized at an even earlier stage of evolution of the Universe, and the intergalactic medium from which the quasars form have the same composition as the interstellar medium. Thus, the age of the heavy elements needed for the formation of Earth-type planets may be comparable with the age of the observable part of the Universe.

The above new data point to the possible existence of planetary systems whose age is close to the age of the oldest objects in the Universe. However, the best evidence that the Earth is not the oldest planet is provided by certain observations as interpreted in the light of the modern theory of stellar evolution. As we have noted before, stars after condensing from the interstellar medium remain in a quasistationary equilibrium for a long time, and the radiant energy losses are balanced by the nuclear reactions in the stellar interior. The length of this phase increases and the luminosity decreases with the decrease of the stellar mass. When the hydrogen has been "burnt up," the stellar nucleus compresses, its temperature increases, and the stellar radius increases. The stars of various masses in which an equilibrium is maintained by thermonuclear fusion reactions (mainly producing helium) constitute the so-called main sequence. Stars which have exhausted their hydrogen supply move from the main sequence to the group of red giants. The duration of the main-sequence phase in the life of a star and the presence of red giants in some group of stars clearly make it possible to find the age of that group. Figure 3 shows the so-called Hertzsprung—Russell diagram for 11 star clusters. The horizontal axis gives the color of the star (the difference between the photographic and the visual stellar magnitudes), and the vertical axis marks the absolute visual stellar magnitude. The envelope on the left is the main sequence curve, and it also plots the color and luminosity distribution of the stars in the youngest of the 11 star clusters, NGC 2362. The vertical axis on the right gives the age corresponding to the duration of the main-sequence phase of a star of a given luminosity. The arrow marks the position of the Sun on the main sequence. The curves branching off the main sequence in the upward right direction plot the color and luminosity distribution of the red giants in each cluster. The branching point evidently gives the age of the cluster. We see from Figure 3 that the branching point of NGC 188 lies below the Sun, which indicates that the age of this cluster is higher than the time that the Sun has so far spent on the main sequence. This conclusion is also borne out by some other data. According to its position in the Galaxy and its velocity relative to the galactic center, the Sun belongs to the disk-type or the intermediate stellar population, which are all characteristic second-generation objects. First-generation objects (the halo subsystem) which formed originally when the galaxies condensed eject gases enriched with heavy elements. These gases are mixed with the leftover interstellar gas of the first condensation, settle to the plane of rotation of the galaxy, and condense into the stars of the disk and the intermediate subsystems—

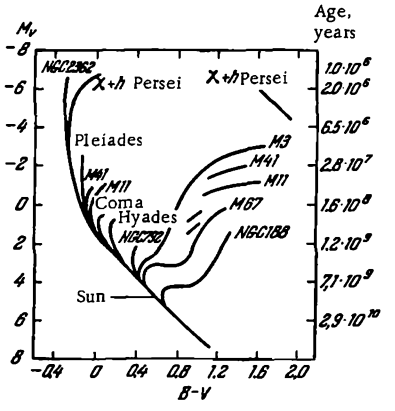


FIGURE 3. Hertzsprung-Russell diagram for some star clusters.

the second-generation stars. Spectroscopic observations of the Sun reveal an abundance of heavy elements characteristic of second-generation objects.

Let us now consider the second component of our dilemma, namely that no signs of activity of supercivilizations have been discovered so far. What is the supporting astrophysical evidence in this respect?

Let us try to estimate the percentage of the currently available astrophysical information out of the total quantity of information which may be contained in the entire electromagnetic spectrum.

Modern astrophysics yields a surprising wealth and variety of information. Optical and radio catalogues list thousands of stars, galaxies, and nebulae. For many of these objects, chemical composition and the physical state of matter are known. The current hypotheses regarding their evolution show a satisfactory fit with the results of statistical analysis of observation data. The observational tools of astronomy have become so effective that radiation sources can be explored at distances of billions of light years. This profusion of data may create the impression that the current hypotheses give a consistent picture of the evolution of the Universe, that almost all the main objects in the Universe have been discovered, and that it only remains to clarify a few minor details.

In my opinion, this is a basically erroneous attitude, although the state of the observational art is such that the structure of the Universe will be elucidated in general outline in the nearest future.

There are numerous examples of outstanding discoveries in astrophysics which were made in recent years only (e.g., the discovery of quasars, the background relic radiation, which accounts for a substantial fraction of the total electromagnetic radiation, molecular generation of the 18-cm hydroxyl line, pulsars). Some of these recently discovered objects may prove to have an immediate bearing on our search for supercivilizations. On the other hand, our knowledge of the quantity and state of solid matter in the Universe is negligible.

As we have noted at the beginning of this chapter, sources of qualitatively new information about cosmic objects may soon become available with the mastering of new frequency regions. What percentage of the entire frequency spectrum have we mastered so far? The search for the main radiation sources in each frequency range is far from providing a complete coverage of all the sources of information, but even this basic problem has not been solved so far. The percentage of the mastered frequencies can therefore be regarded as an upper bound estimate of the available quantity of information. It is in this sense that we should interpret the concept "mastered frequency range."

A frequency range is said to have been mastered if more than 30% of the total sky area has been scanned for sources at a given wavelength, and more than 100 cosmic objects have been discovered as a result of this search. We have to distinguish between two cases:

1. The search for objects emitting a wide spectrum of frequencies (spectrum width $\Delta\nu \sim \nu$).

2. The search for objects emitting in narrow spectral lines.

The second problem is clearly incomparably more complex than the first, since it involves coverage of the entire electromagnetic spectrum

with narrow-band filters. Roughly speaking, the number of measurements required in case 2 is a factor of $\frac{\nu}{\Delta\nu}$ greater than in case 1. For the radio lines of the interstellar hydroxyl OH at 18-cm wavelength we have for some objects $\frac{\nu}{\Delta\nu} = 3 \cdot 10^6$.

This narrow-band scanning is of the greatest importance both for astrophysics and for the search for civilizations. So far, however, no narrow-band survey of the sky has been carried out either in the optical or the radio spectrum (the only possible exception is the complete survey of the sky in the interstellar hydrogen line $\lambda = 21$ cm in a band of about 1 MHz with spectral resolution of about 10 kHz). The percentage of the available information on pure monochromatic sources is therefore still exceedingly small.

The search for wide-band sources is a much simpler problem. The number of mastered frequency ranges (e.g., octaves) for these sources is determined by the expression $I \propto \ln \frac{\nu_2}{\nu_1}$, where ν_1 and ν_2 are the minimum and the maximum frequency of the survey. The percentage of the mastered frequencies is clearly given by

$$x = \frac{\ln \left(\frac{\nu_2}{\nu_1}\right)_{\text{rad}} + \ln \left(\frac{\nu_2}{\nu_1}\right)_{\text{opt}}}{\ln \frac{\nu_2}{\nu_1}}, \qquad (1.3)$$

where $\left(\frac{\nu_2}{\nu_1}\right)_{\text{rad}}$ is the maximum to minimum survey frequency in the radio spectrum, $\left(\frac{\nu_2}{\nu_1}\right)_{\text{opt}}$ ditto in the optical spectrum, and the ratio $\frac{\nu_2}{\nu_1}$ in the denominator is determined by the maximum and the minimum frequencies of astronomical surveys in future.

At present, radio surveys have been conducted at frequencies between 40 and 400 MHz, so that $\left(\frac{\nu_2}{\nu_1}\right)_{\text{rad}} = 10$.

In the optical spectrum, photographs and observations of individual sources covered the range from 3000 to 6000 Å, i.e., $\left(\frac{\nu_2}{\nu_1}\right)_{\text{opt}} = 2$.

In the other frequency ranges, there are only isolated observations of small sky areas, which constitute a negligible percentage of the entire quantity information.

What is the value of the denominator in (1.3)? The low-frequency limit of astrophysical observations has been fixed with fair certainty. The minimum frequency is $\nu_1 \sim 1$ MHz, since at lower frequencies the interstellar medium is opaque and only objects very close to our planetary system can be observed.

The high-frequency limit is more difficult to determine, and it apparently linked with the quantum nature of electromagnetic radiation. As the frequency increases, the energy of each detected quantum becomes higher. Now, as the energy resources of astronomical objects are limited, the number of quanta reaching the detector decreases as the quantum energy increases. A more detailed estimate of the frequency ν_2 will be given

later on. For the time being we take $\nu_2 \sim 10^{18}$ Hz (wavelength 3 Å). Then $\frac{\nu_2}{\nu_1} = 10^{12}$, and the percentage of mastered frequencies is

$$x = \frac{\lg 10 + \lg 2}{12} = 11\%, \tag{1.4}$$

i.e., even in the relatively easy search for wide-frequency bands, we have so far mastered a low percentage of the total information available. Note that of the 89% of the missing information, 42% falls between 10^9 and 10^{14} Hz (centimeter, millimeter, submillimeter, and infrared waves) and 25% between 10^{15} and 10^{18} Hz (ultraviolet radiation and X-rays).

The limits ν_1 and ν_2 of the entire electromagnetic spectrum are fixed with considerable uncertainty. We have probably underestimated its width, so that the 11% is an overestimate.

Let us now estimate the number of cosmic sources which can be discovered in a given electromagnetic frequency range. As we have noted before, the sensitivity of some radiation detectors has now almost reached the physical limit determined by the quantum nature of the electromagnetic radiation and the background of cosmic radiation. Therefore, the success of a search for sources of small angular dimensions will depend on the number of quanta per unit detector surface area and the possibility of resolving the various sources.

On the long-wave side the number of sources is limited by the angular resolution of the antenna. The number of antenna beam widths accommodated by the celestial sphere is

$$N_1 \leqslant \frac{A_\nu}{\lambda^2} = \frac{A_\nu \nu^2}{c^2}, \tag{1.5}$$

where A_ν is the effective collecting area of the telescope. In the radio spectrum, the best antennas have $A_\nu \propto \lambda^2$. This is so because the relative precision with which a reflecting surface can be manufactured is approximately constant, i.e., $\frac{\varepsilon}{D} \sim$ const, where D is the reflector diameter, and ε is the mean error surface; for a reflector to be effective in a given frequency range, we should have $\varepsilon \leqslant 0.1\lambda$. Thus, in the radio range, the maximum number of distinguishable sources N_1 is independent of wavelength.

A survey of the hundred brightest sources in every frequency range clearly does not require antennas of maximum capacity. Nevertheless, taking $N_1 \sim 100$, we should change the effective area $A_\nu \propto \lambda^2$ on passing from one frequency range to another.

Relation (1.5) leads to an important conclusion. When working with the instrument of maximum capacity and when surveying different frequency regions for a constant number of the brightest sources, the expected quantity of information is proportional to $\ln \frac{\nu_2}{\nu_1}$ and the above estimates based on (1.3) remain valid.

For short-wave observations (X-ray and gamma-ray frequencies), we can work with equipment counting every single incoming quantum and faithfully indicating the direction from which it arrived. The number of sources that can be discovered in a time τ therefore cannot exceed the number of quanta from these sources which reached the detector,

$$N_2 \leqslant \frac{c\rho A_\nu \tau}{h\nu}. \tag{1.6}$$

Here ρ is the total density of electromagnetic radiation in a given frequency range in unit volume from all the sources. According to measurements at wavelengths shorter than the optical spectrum $\rho < 10^{-12}$ erg/cm^3. The parameter A_ν (e.g., the cross section of the gamma counters) hardly changes with wavelength in this case, and therefore N_2 diminishes as the frequency increases. Clearly, the frequency at which $N_1 \sim N_2$ is that particular ν_2 above which only a negligible fraction of information is contained. (A_ν cannot be increased with increasing frequency because of formidable technical difficulties.)

Thus, equating (1.5) and (1.6) and assuming A_ν to be of the same order, we find

$$\nu_2 \sim \left(\frac{\rho c^3 \tau}{h}\right)^{1/3}. \qquad (1.7)$$

Because of the weak dependence of ν_2 on the particular values of the parameters, we may take $\rho \leqslant 10^{-12}$ erg/cm^2, survey time $\tau \sim 1$ year $\sim 3 \cdot 10^7$ sec, and this gives $\nu_2 \leqslant 5 \cdot 10^{17}$ Hz.

Let us briefly reiterate the conclusions which follow from the above discussion: despite the great advances in astrophysics, our information is still insufficient to disprove the possible existence of supercivilizations by arguing that so far no signs of their activity have been observed. At a later stage we will consider the possibility that some of the already known objects (e.g., quasars) are in fact products of activity of supercivilizations. On the other hand, the astrophysical data firmly indicate the existence of planetary systems much older than the solar system. This provides justification for setting up a detailed program of search for extraterrestrial civilizations.

We have considered some of the astrophysical aspects of the fundamental dilemma (C) and our conclusion is that the entire dilemma is most probably a product of insufficient knowledge on our part. If this is indeed so, we must try to establish what astrophysical signs the activity of supercivilizations can be expected to produce. This problem probably can be solved by analyzing some general features of the development of civilizations over cosmogonic periodics. It should be clearly understood that our knowledge in this field is pitiful. On the other hand, we will not be able to go any further without making some basic assumptions. There is no doubt that the laws governing any field of activity of our civilization can and should be formalized and systematized to a certain extent. This approach will probably prove helpful in our analysis also. Some general considerations on this subject are given in the next section.*

§4. CIVILIZATIONS AND THE MAIN FEATURES OF THEIR DEVELOPMENT

We are primarily concerned with the highest level of development and the general trend of activity of civilizations which we can expect in the initial phases of the search program. Once these preliminary points are settled, we will be able to reach certain conclusions regarding the

* Also see Chapters V and VI.

observable signs of this activity on cosmic scales and to analyze the possibilities of detection of these signs with modern means.

The main factor which has been firmly and reliably established by modern astrophysics is the universality of all the fundamental laws of nature everywhere in the observable part of the Universe and over the entire period of time covered by the evolutionary scale. We may therefore assume with fair likelihood that the physical laws known to us are also known to supercivilizations. The knowledge of supercivilizations clearly may cover a much wider gamut of physical laws, but the sum total of their knowledge will contain as a subset all that we know. Moreover, the present level of our technical and scientific knowledge is apparently an unavoidable and necessary step in the early development of any technical civilization. We can thus try to formulate in crude terms some general concepts applicable to all extraterrestrial civilizations.

A functional definition of a civilization is highly important for future use. A detailed discussion of the functional definition of life, originally proposed by Lyapunov /9/, is given in /1/ (pp. 125—132):

a highly stable state of matter capable of developing survival reactions using data coded by the states of the individual molecules (D).

This definition adequately conveys the main content of the concept, but in our opinion it has one fundamental shortcoming: it does not mention the general laws and features governing the conception, development, and evolution of various life forms. The life of any individual apparently can be considered as a stochastic process governed by its interactions with the environment and the state of the live object at any given time. The evolution of the species in this case is regarded as a certain statistical law which emerges from the growth and development of the individual organisms. An obvious outcome of evolution is a steady accumulation of information and its adaptation to practical applications. Therefore, it seems to us that the main statistical trend in the development of living organisms is the tendency to gain the maximum quantity of information about the environment and about the organism itself (E).

For the lower life forms, this trend is dictated by natural selection. This also seems to be the only stimulus for the development of the higher forms of civilization.

The distinctive feature of the higher life forms is their ability to undertake an abstract analysis of the acquired information. Systems of living organisms begin to play an increasingly important role as the life forms develop. However, we can hardly fix at this stage the exact number of organisms and the structure of a high-level civilization. Thus, bypassing the above definition of life, we can offer the following functional definition of a high-level civilization:

a highly stable state of matter capable of acquisition, abstract analysis, and application of information for the purpose of extracting the maximum quantity of information about the environment and itself and developing survival reactions (F).

There is no need to include a specific coding mechanism in this general definition. Information about environment and self includes all data about

animate and inanimate nature (including civilization), science, technology, culture, art. (There are probably other, hitherto unimaginable fields which also should be included in this category.)

If we accept definition F, the principal parameters characterizing the degree and the character of development of a civilization are the quantity of information and the rate of accumulation of new information (e.g., the time to double the sum total of knowledge). Within the framework of modern concepts (and here we have to differ with von Hoerner /2/, p. 278), it seems to us that definition F allows for an unlimited development of civilizations. Von Hoerner's principal hypotheses regarding the limit of development of civilizations include 1) total destruction of all life, 2) destruction of intelligent life, 3) degeneration, 4) loss of interest. These suicidal factors apparently acquire great significance for every civilization at a certain stage of development, but there is no proof that they are fundamentally unavoidable in every case for all civilizations. The only reason for a civilization to stop developing in the light of definition F is the existence of a finite quantity of information in all the fields. This, however, seems to be a most unlikely propostion.

A highly important aspect for the search program is that the quantity of information in certain fields is finite (this, naturally, does not imply that the total quantity of information is finite). One of these fields with a finite quantity of information is possibly space science at its present level. To make this point, consider the following example. We have already mentioned that the modern methods of astrophysics enable us to study various objects in the Universe billions of light years distant from the Sun. For these distances, the very concepts of length and time of light propagation are not single-valued, and they significantly depend on the particular model of the Universe used. The main method for estimating the distances of extremely far objects is the determination of the change in the wavelength of the emitted spectral lines (relative to the laboratory wavelengths), i.e., the red shift z. As we have mentioned before, spectra of sources with $z \sim 2$ have now been obtained. At the same time, radio sources with the weakest observable continuous spectra may have a substantially higher z. Were it not for absorption and scattering of electromagnetic radiation in the intergalactic medium, the largest modern radio telescopes could detect quasar-type objects with $z \sim 30$, and the projected radio telescopes could in principle advance this limit even farther. However, calculations and statistical analysis of radio observations show that this is not so.

The main factor preventing the effective observation of these ultra-distant objects is apparently the scattering of electromagnetic radiation by free electrons in the intergalactic and galactic medium. This effect, as demonstrated in /10/, fixes $z \sim 5$ as the most probable maximum distance at which radio sources are still observable (this is the value obtained for a positive curvature model with $q_0 = 1$, $H_0 = 300$ km/sec $\cdot 10^9$ light years, and the present-day density of the intergalactic medium $\rho_0 \sim 4 \cdot 10^{-29}$ g/cm^3). Although no direct determinations of the density of the intergalactic medium are possible at this stage, a statistical study of radio sources shows that the number of weak sources is less than what could be expected without scattering. The theoretical result which points to the existence of the

maximum accessible distance thus appears to have a certain experimental justification.*

The sphere characterized by maximum z contains a finite quantity of matter, i.e., a finite number of cosmic objects. Since the structure of celestial bodies is described by the same general laws in different parts of the Universe, it is quite probable that the principal properties of all these objects will require only a finite time to study.

In all likelihood, many of the principal laws of nature will be established within the next decade in view of the current tendency of astrophysical research (A). Thus, the information concerned with space science has an objectively finite limit, and there is a definite possibility that the supercivilizations may lose all interest in this science. This, in particular, may resolve our dilemma (C) — there is no universal civilization because the highly developed civilizations have lost all interest in space research. By space research we naturally mean research in the modern astrophysical sense. There may be certain directions associated with space science of which we are not aware at present (e.g., problems connected with universal physical constants) and in which there is promise of an unlimited quantity of information.

We should again stress that the problem of acquiring a complete quantitative knowledge of the laws of the Universe is essentially simplified by the inherent similarity of the celestial objects in various parts of the Universe, as is evident from the currently available astronomical data. Civilizations themselves are apparently the only type of objects which do not follow this law of uniformity. Therefore, to ensure a maximum rate of acquisition of new knowledge, the best way is to strive toward information exchange between civilizations. In the light of modern ideas, exchange of information through space is most effectively accomplished by means of electromagnetic radiation. It is moreover clear that the most general factor associated with the activity of supercivilizations is the use of mass and energy on a gigantic scale.

In trying to distinguish between the activity of civilizations and the effects of natural processes in the Universe, we should apparently be guided also by the above definition of the civilization.

We cannot give any sound quantitative estimate of the maximum level of development of supercivilizations. However, since there is a very good chance of our mastering the entire electromagnetic spectrum and thus markedly increasing the sum total of our astronomical knowledge, we hope that this estimate will come within our reach some time in the future.

The present-day astrophysical data do not impose any limit on the possible development of supercivilizations, which in principle may reach fantastically high levels. It may even be argued that the expansion of the observable part of the Universe may conceivably be a result of some intelligent activity of a supercivilization. According to the modern models of the expanding Universe, all matter was in a superdense state some 10 billion years ago. Does this preclude the continuous existence of civilizations at earlier stages of evolution, 20, 100, and 1000 billion years ago, or is there a possibility that they survived the instant when the Universe was

* Another reason which interferes with the observation of distant sources is the absorption of their radiation by nearer sources. Already for $z \approx 2$, the probability that the line of sight intercepts more than one object is close to 1.

in the superdense state? The age of the oldest civilizations can be reliably fixed at a few billion years only when we shall have firmly established that prior to the expansion the conditions in the Universe were adverse to the inception and development of life.

Can we describe in general outline the development of a civilization over cosmogonic periods? We know that many of the fundamental parameters characterizing the development of the Earth civilization grow exponentially (see Chapter V). The time to double the scientific and technical information is about 10 years, the time to double the power resources, the raw material reserves, and the population is about 25 years. Extrapolation of the current rates of growth of our society to the nearest future therefore leads to curious paradoxes.

In a book by a group of outstanding American authorities on thermonuclear reactions /11/, the authors call our attention to the fact that the quantity of energy that can be generated on the Earth is not very high. There is a definite upper limit to it. The Earth absorbs (and re-emits) $5 \cdot 10^{23}$ erg of solar radiation each second. To avoid drastic changes in the Earth climate, the energy output of artificial installations on the Earth must be limited approximately to one percent of this quantity. Assuming a figure of $4 \cdot 10^{19}$ erg/sec for the current power output and an annual growth of 4 percent, the authors show that the upper limit will be reached in 125 years! This limit can be slightly stretched if we directly harness the solar radiation. To this end, however, a considerable part of the Earth's surface will have to be covered with solar energy converters, a not very likely prospect.

Thermodynamic considerations show that this is indeed a fundamental difficulty. After all, the entire expended energy is inevitably converted into heat. And what then? Two solutions can be envisaged: either the power output is maintained strictly constant after the allowed 125 years of growth, or all the forms of human activity involving large energy requirements (industrial complexes and large-scale scientific experiments) should be moved into outer space. The first alternative is entirely unacceptable, since it virtually means that all further development is stopped. The second alternative, on the other hand, appears quite likely even at the present stage of development.

A similar conclusion regarding the inevitable expansion into outer space also emerges from an examination of other characteristics of human activity (population explosion, chemical and radioactive contamination of the ocean, insufficient open space on the Earth, exhaustion of nuclear fuel resources, shrinkage of the biosphere, etc.). Power difficulties, however, will probably prove the dominant motivating factor. If a certain parameter P increases a factor of α annually, P_0 will increase in t years to $P = P_0 \alpha^t$, whence

$$t = \frac{\lg (P/P_0)}{\lg \alpha} \text{ years.} \tag{1.8}$$

The above estimate of 125 years was obtained using this relation. If the growth rate $\alpha = 1.04$ is maintained after the critical period, the human power output will exceed the quantity of incident solar radiation after 240 years, after 800 years the total energy radiated by the Sun will be exceeded, and after 1500 years we will exceed the total radiation output of the entire Galaxy!

The population also grows exponentially, and possibly even faster, so that there will be a steady pressure to maintain the exponential growth of the other parameters. So far, our civilization has used up about 10^{17} g of mass. Assuming the same annual growth rate, the figure will reach 10^{51} g in 2000 years, which is equivalent to the mass of more than ten million galaxies! The quantity of information currently increases at a rate of 10% annually; extrapolating for this rate of growth, we obtain an increase by a factor of 10^{80} in 2000 years, so that the quantity of information by then will significantly exceed the total number of atoms in the Universe (about 10^{80}). Such a quantity of information in principle cannot be stored or remembered! We thus reach the inevitable conclusion: the current exponential growth constitutes a transient phase in the development of the civilization and it will be unavoidably restrained by natural factors.

Indeed, assuming a mean density of ρ for some space medium that the civilization has set forth to harness, we see that, even advancing at the velocity of light, it will be able to harness, after some time t, mass at a rate not exceeding

$$\frac{dM}{dt} \leqslant 4\pi (ct)^2 \rho c, \tag{1.9}$$

and energy at a rate not exceeding

$$\frac{dE}{dt} \leqslant 4\pi (ct)^2 \rho c^3. \tag{1.10}$$

Hence it follows that the mass and energy requirements (and therefore the growth of information, whose material carriers are mass and energy) may increase exponentially only for a limited time, whereas an unlimited growth may not be faster than . For our civilization, as we have seen, the duration of the future exponential growth phase can be estimated at about 1000 years. And what then? Since the development of power resources on the Earth (and, in general, in any finite volume) is limited by thermodynamic considerations (the overheating effect that we have mentioned before), future economic growth after some 100 — 200 years will probably push humanity into outer space! This, in our opinion, is the objective tendency and the main task of space exploration at this stage.

Should not the nonexponential growth be interpreted as a sign of a decaying civilization? In our opinion, even a linear growth of information indicates a viable civilization. Indeed, a constant rate of acquisition of information signifies that a constant quantity of new, highly significant and highly valuable data is acquired every year. This in no way obstructs the main tendencies in the development of civilizations. The so-called "feedback effect" will apparently constitute a decisive factor for further development of our civilization. Everything depends on whether supercivilizations exist or not. If the answer is in the affirmative, reception and assimilation of information from supercivilizations may play a leading role in future development. This learning stage may lead to a rapid jump of the civilization to the highest level. If we assume that every civilization at a certain stage of its development passes through such a learning stage, we conclude that there will be virtually no civilizations in an intermediate stage of development or in a stage close to ours.

The second possibility — total absence of supercivilizations — will apparently necessitate a complete revision of our current ideas of unlimited growth and development.

§5. THE SEARCH FOR SIGNS OF ACTIVITY OF SUPERCIVILIZATIONS

The general considerations of the previous sections lead to certain conclusions regarding the types of activity of supercivilizations which can be detected at the present level of development.

The most general parameters of this activity are apparently ultra-powerful energy sources, harnessing of enormous solid masses, and transmission of large quantities of information of different kinds through space. In this section we will consider the first two parameters which are a prerequisite for any activity of a supercivilization.

Energy sources

As we have noted before, the present-day astrophysical observations do not provide any indication of the existence of an upper limit for the energy output of a supercivilization. This limit, however, will probably emerge when we have covered the entire electromagnetic spectrum, from 10^6 to 10^{18} Hz. This interesting conclusion follows from basic thermodynamic considerations: the entire energy expended by a supercivilization is inevitably converted to heat. This thermal energy cannot accumulate indefinitely inside a closed volume, to avoid critical overheating. The only way in which this heat can be dissipated is by radiation into outer space. Any power system thus inevitably involves eventual radiation of its entire power output in the form of heat into space. If the efficiency of these systems is very high, the spectrum and the surface brightness of the radiating body should correspond to the blackbody spectrum at a temperature equal to the effective temperature of all the forms of electromagnetic radiation received from outer space (the equilibrium temperature in the intergalactic medium is around 3°K). It is quite probable, however, that the efficiency of these power systems is less than 100% (there can be various operational reasons for this). The resulting emission spectrum is more complex. It is difficult to predict the specific features of sources of this kind. The only reasonable thing to do at this stage is to concentrate on radiation sources with maximum bolometric power. Quasars are the only known objects which fall under this category.

Let us briefly describe the main regular features established for these remarkable objects /8/.

The radio emission of quasars was discovered more than 10 years ago, but the widespread interest in these objects was aroused only recently. In 1960 — 1962, following a substantial improvement in the directivity on radio observations, it was established that some radio sources have the same coordinates as star-like optical objects. Prior to that time, the consensus of opinion had been that most radio sources are identifiable

with large galaxies. It thus appeared that a new class of stars with anomalously powerful radio emission had been discovered in the Galaxy. Further observations, however, proved this hypothesis to be wrong. The observational data indicated that these were an entirely new type of extragalactic object, of which nothing had been known before. (Note that this again stresses the need to cover the entire electromagnetic spectrum in our observations.)

When observed through optical telescopes, the quasars appear as star-like objects in the sense that the apparent angular diameter is substantially less than the resolution limit of the astronomical optics (fractions of an angular second). Near some of the quasars, nebulous filaments are observed, which may be irregular in shape or follow a general radial direction from the star, reminiscent of ejected gases. Figure 4 is a photograph of one of the nearest and brightest quasars, the radio source 3C 273, with a noticeable ejection on top right. The ejection is no wider than 1" — 2", it begins at a distance of 11" from the star and terminates at a distance of 20". Ejections and filaments are also observed near the quasars 3C 48, 3C 196, and 3C 279.

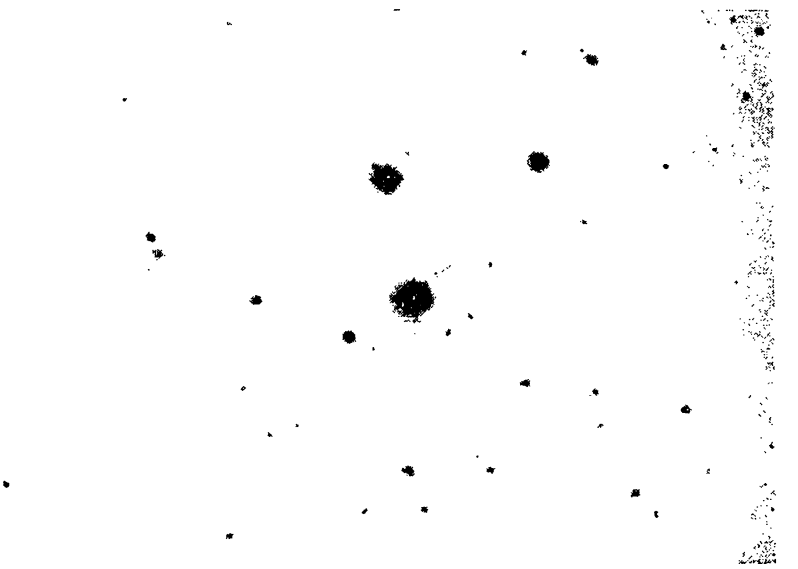


FIGURE 4. The quasar 3C 273.

One of the most remarkable features of the optical spectra of quasars is the exceptionally strong red shift of the spectral lines. The red shift varies from 0.158 (for the nearest quasars 3C 273) to values corresponding to a three-fold change in wavelength (3C 9, see Figure 2). This unusually high red shift, if interpreted as the result of the expansion of the Universe, points to tremendous distances and fantastic luminosities of these objects.

Because of the high red shift, the optical spectrum contains some lines which normally lie in the ultraviolet in laboratory spectra. For ordinary stars and galaxies, this spectral region is inaccessible to observations from the Earth, as the atmosphere is opaque to wavelengths shorter

than 3000 Å. The spectra of quasars have by now been studied down to 1000 Å, and some spectra actually gave the profile of the L_α hydrogen line — the strongest line of most cosmic objects. Table 1.1 lists the elements and the ionization stages discovered in the spectra of quasars. The first column gives the elements in the order of increasing atomic number, the second column itemizes the observed ionization stages. The missing lines are apparently those of elements which occur in small quantities, in accordance with their normal abundance, or which normally do not have bright lines in the observed part of the spectrum. The logarithm of the normal abundance (by number of atoms) is given in the last column of the table.

TABLE 1.1

Element	Ionization	Abundance	Element	Ionization	Abundance
H	I	12.0	P	—	5.53
He	II	11.16	S	II	7.22
Li	—	3.0	Cl	—	5.4
Be	—	2.4	Ar	IV	6.62
B	—	2.8	K	—	4.88
C	II, III, IV	8.48	Ca	II	6.22
N	IV, V	7.96	Sc	—	2.91
O	I, II, III	8.83	Ti	III	4.82
F	—	5.4	V	—	3.78
Ne	III, V	8.44	Cr	III	5.38
Na	—	6.22	Mn	II, III	5.10
Mg	II, V	7.46	Fe	II	6.90
Al	II, III	6.28	Co	II	4.72
Si	II, III, IV	7.47	Ni	II	5.93

The conditions of excitation of spectral lines in quasars are apparently highly variable. Some quasars show mainly emission lines, most of which can be identified with the spectra of certain elements. Figure 5 is a microphotometric tracing of the spectrum of 3C 273 /12/. In addition to emission lines, the spectrum shows wide emission bands of uncertain origin. Figure 6 is the profile of the H_β line in the spectrum of this quasar /13/. Some features of the line profile show distinct signs of a shift relative to the line center. The Doppler velocities corresponding to this shift are as high as a few thousands of kilometers per second. Some quasars have a rich spectrum which also contains absorption lines and bands (e.g., 3C 191). Some quasars (e.g., 3C 682) do not show any lines at all.

The continuous optical spectrum of quasars also shows a number of characteristic features. The energy distribution in the quasar spectra is markedly different from the energy distribution in the stellar spectra. Quasars can thus be readily identified in large-scale measurements of star color with light filters. The energy distribution in the optical spectra of quasars is adequately described by a power function $F_\nu \propto \nu^{-\alpha}$, and a probable mechanism is therefore emission or scattering of radiation by relativistic electrons.

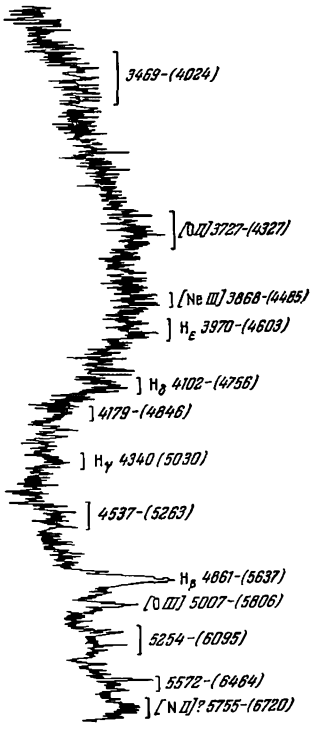


FIGURE 5. The microphotometric tracing of the spectrum of the quasar 3C 273.

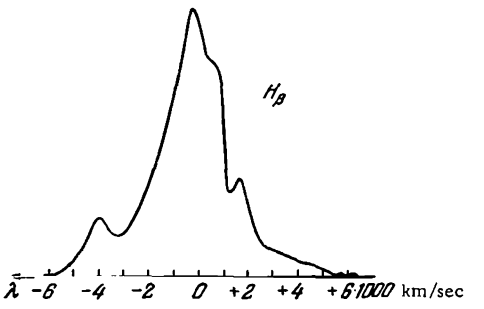


FIGURE 6. The profile of the H_β line of the quasar 3C 273.

Figure 7 is a plot of the optical colors obtained with three filters, U (λ 3600 Å), B (λ 4400 Å), and V (λ 5500 Å). The horizontal axis gives the $B - V$ difference, and the vertical axis the $U - B$ difference for the same object. The lower curve is the locus corresponding to the main-sequence stars, and the top line is the power energy spectrum. The quasar region is cross hatched.

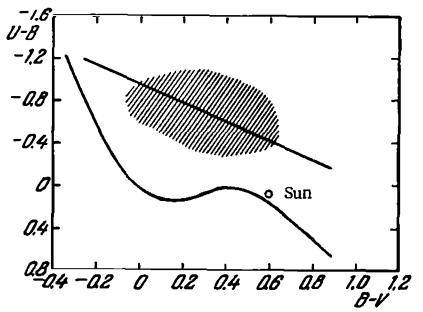


FIGURE 7. The colors of main-sequence stars and quasars.

One of the most puzzling properties of quasars are the variations of their intensity. Prior to the discovery of quasars, extragalactic astronomy was generally assumed to deal with highly stable sources. The brightness of galaxies remains constant over billions of years (except for the brief supernova

explosions). And yet, the observations of the first quasars have shown that their luminosity is significantly variable. Using old photographs of the sky, the astronomers managed to reconstruct the light curves of these objects over a relatively long period. Figure 8 shows the smoothed light curve of 3C 273 for the period 1888–1963 /14/. The mean light variation period of this source is about 9 years. The mean photographic magnitude of 3C 273 decreases according to the equation

$$m_{pq} = 12^m.47 + 3^m.67\,(T - 1900),$$
$$\pm 0.08 \qquad \pm 0.47$$

(where T is the year of observation), which gives 300 years for an exponential decrease of brightness to $1/e$ /15/. Faster brightness fluctuations, whose statistical character is still unclear, have also been observed.

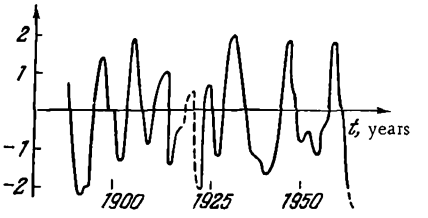


FIGURE 8. The smoothed light curve of the quasar 3C 273 corrected for the secular decrease in brightness.

Figure 9 plots the results of photographic and photoelectric measurements of the stellar magnitude with a B filter for the quasar 3C 446 /16/. Occasionally, the brightness of this object changes by as much as a factor of 2 in 24 hr! This rapid variation of brightness provides a direct estimate of the size of the emitting region – less than one light day ($< 3 \cdot 10^{15}$ cm), i.e., much less than the size of a galaxy (tens of thousands of light years) and probably even less than the size of the solar system: the diameter of the orbit of Pluto is 0.5 of a light day.

Both the continuous and the line spectrum of quasars apparently change (the changes cover line widths, line intensities, and wavelengths /17/). The correlation between these variations has been hardly studied.

Some quasars show a considerable linear polarization of the optical radiation. The same quasar 3C 446 has a maximum difference of $0'''.2$ between the intensity of the perpendicular polarization components. The degree of polarization and the position angle apparently change with time. The polarization of the infrared radiation at $\lambda = 1.6\mu$ for the quasar 3C 273 reaches 40%.

Let us now consider some fundamental results of radio observations of quasars. The angular resolution of the modern radio telescopes can be made as high as $0''.001$ (by using interferometric techniques, observing the diffraction pattern during lunar occultation of radio sources, and

studying the radio flux fluctuations associated with radio wave propagation in an inhomogeneous interplanetary medium), and this is considerably higher than the resolution attainable with optical telescopes for the same objects.

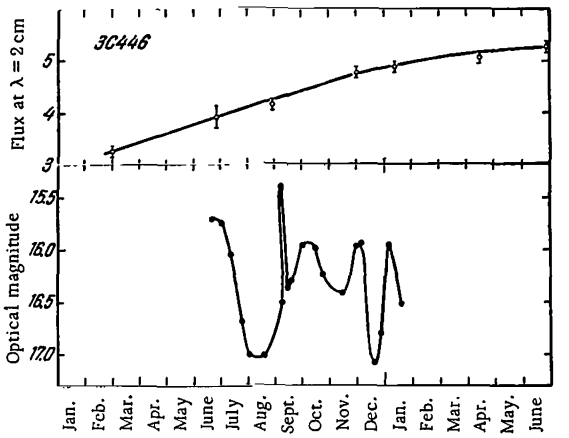


FIGURE 9. Light curves of the quasar 3C 446.

However, not much information has been obtained so far by the new radio methods. The observations of 3C 273 (the best studied quasar) revealed the existence of two sources: source A corresponding to a luminous ejection on the photograph of this object, and source B which adequately coincides in position with the quasar itself. Figure 10 shows the radio spectra of components A and B, which are markedly different /18/.

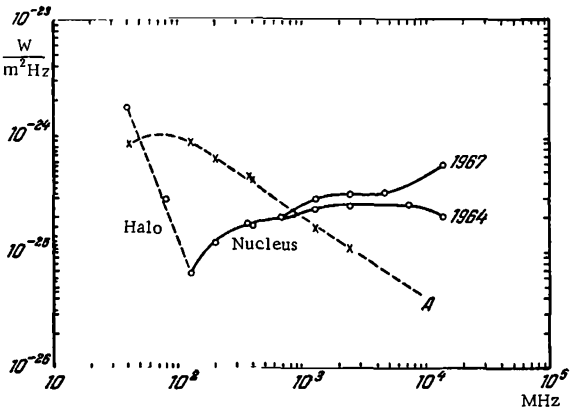


FIGURE 10. The spectrum of the components of 3C 273.

Source *A* is elongated along the optical ejection and grows brighter at the outermost end, where its angular dimensions are 5" × 1".5. Source *B* in its turn consists of a spherical halo some 6" in diameter and a central nucleus /19/. Radio-interferometric observations reveal that most of the energy is radiated from a region not exceeding 0".002 /20/.

The spectra of quasars often deviate from the normal power function, and this probably suggests a complex structure or a variety of emission processes. The most interesting properties are those of sources with peculiar features in the short-wave part of the radio spectrum. Figure 11 shows the spectrum of 3C 279; like the spectrum of 3C 273B, the radio flux shows a tendency to increase toward shorter wavelengths.

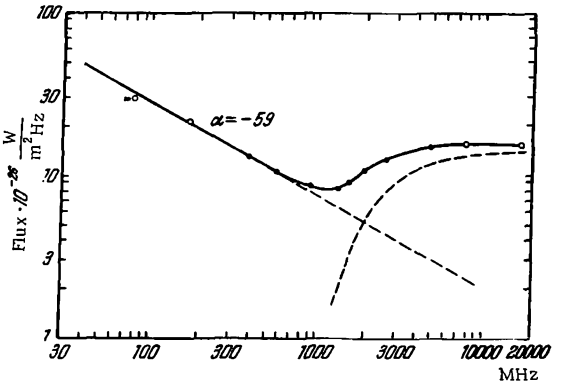


FIGURE 11. The spectrum of 3C 279.

The radio emission of these objects is generally variable. Figures 12 and 13 plot the time variation of the radio flux from these two sources at various wavelengths /21/. Particularly strong and rapid variations are observed in the millimeter range. In 1966, a decision was taken to launch an international program of systematic observations of selected objects in the entire electromagnetic spectrum in order to study the variability of quasars. The sources 3C 273, 3C 279, 3C 345, CTA-102, and others were chosen for this purpose. The list also included the source 3C 84, which is a nucleus of the anomalous galaxy NGC 1275. The properties of this source have much in common with the properties of quasars. Detailed observations also reveal a deep-running analogy.

No individual radio lines from quasars have been observed thus far, since every quasar requires special receiving equipment adjusted to its red shift.

The brief description of the observational data shows that our information about quasars is highly deficient even in the well-mastered frequency ranges and for the brightest sources. It is quite probable that some of these sources have an exceptionally strong radiation in the intermediate spectral region (between the radio and the optical spectra).

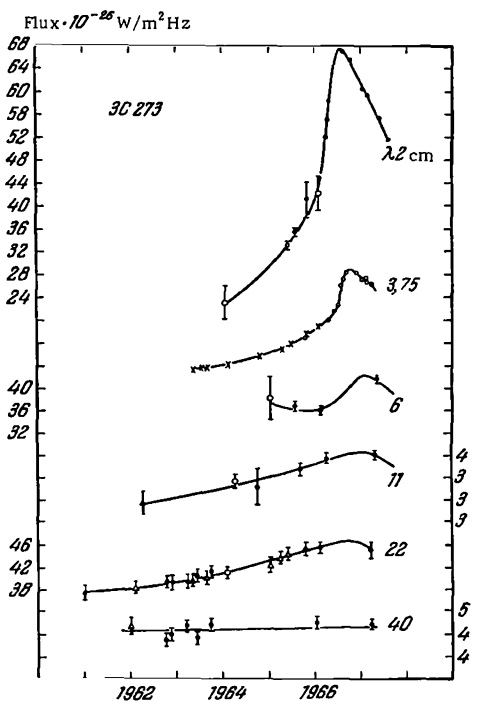


FIGURE 12. Variation of radio flux, degree of polarization, and position angle for the quasar 3C 273 at 8000 MHz.

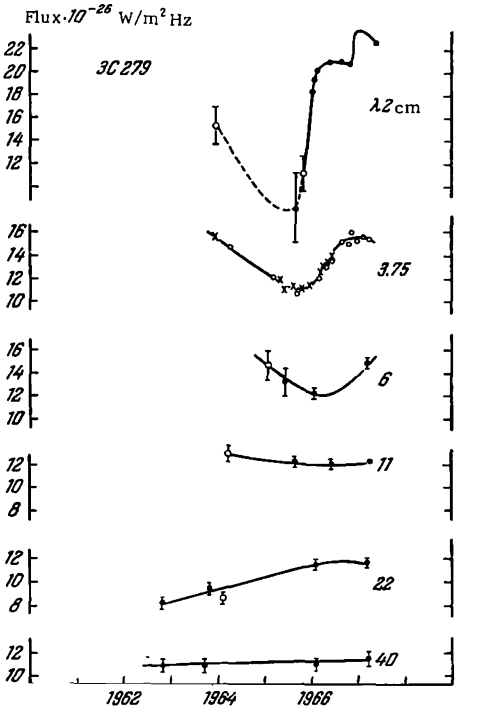


FIGURE 13. Variation of the radio flux from the quasar 3C 279 at 8000 MHz.

The bulk of the energy of 3C 273, for instance, is definitely known to be radiated in this range. Figure 14 shows the combined spectrum of 3C 273 B based on both radio and optical observations, plus the new measurements in the millimeter and the infrared spectra /23/. The steeper short wave curve is based on the 1964 measurements, and the gentler curve is the result of 1966 measurements. The spectrum of quasars probably extends far into the ultraviolet and the X-ray region. Recently, 3C 273 was apparently found to emit at 1 — 10 Å /24/.

The total bolometric luminosity of the quasars is unusually high. The total flux emitted in the infrared and in the submillimeter region by 3C 273 B reaches $4 \cdot 10^{-12}$ W/m^2. Since the distance to the source is $< 1.5 \cdot 10^{27}$ cm, the total energy radiated in this range is about 10^{47} erg/sec. The energy emitted in the optical spectrum is $^1/_{10}$ of this value, and that in the radio spectrum $^1/_{100}$ of this value. Thus, there apparently exist quasars which are basically infrared sources /25/. The bolometric power of quasar 3C 273 is thousands of times greater than the corresponding power of the giant galaxies.

Studies of the spatial distribution of quasars revealed still another remarkable feature: quasars never occur in clusters or near individual galaxies /26/.

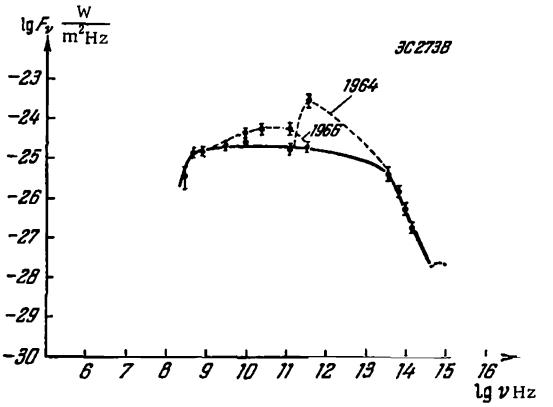


FIGURE 14. The spectrum of 3C 273B according to radio, infrared, and optical measurements.

Very interesting conclusions emerge from statistical studies of the distribution of quasars according to the observed radiation flux. These statistics reflect the line-of-sight distribution of sources. If we allow for the time of propagation of the radiation, this distribution can be taken as characterizing the quasar number and power at various stages of evolution of the Universe. Figure 15 plots the function $N(F_\nu)$, i.e., the number of all radio sources brighter than a given flux F_ν vs. the flux. The curve is based on the observations of all the radio sources at 178 MHz up to a maximum flux of $5 \cdot 10^{28}$ W/m^2·Hz /26/.

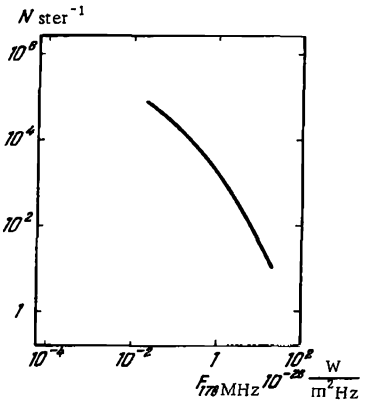


FIGURE 15. The number of sources brighter than a given flux vs. the flux value at 178 MHz.

Theoretically, a uniform distribution of radio sources in a Euclidean space without expansion gives $N(F_\nu) \propto F_\nu^{-3/2}$.

It follows from the theory that the red shift associated with the expansion of the Universe should lead to a more gentle dependence. Observations, on the other hand, give a steeper dependence:

$$N(F_\nu) \propto F_\nu^{-1.8}.$$

Recently it has been established that if all the sources are divided into two groups — quasars and radio galaxies — each class will have its own distribution function $N(F_\nu)$. For radio galaxies $N(F_\nu) \propto F_\nu^{-3/2}$, and for quasars $N(F_\nu) \propto F_\nu^{-2.2}$ /27/.

The possible reason for this steep distribution is the rapid evolution of the radio sources in an expanding Universe (either a decrease in the number of sources in every bounded volume with the expansion of that volume, or

a decrease in the brightness of the source, or both). There is a possibility that only quasars are characterized by this exceptionally fast evolution.

If the division of sources into two groups is justified, the number of quasars among sources with fluxes of 10^{-26} W/m^2·Hz is comparable with the number of other sources (from observations at 178 MHz). The curve $N(F_\nu)$ then also shows a marked saturation at low fluxes. The low brightness temperature of the extragalactic background (about 20°K at the same frequency /26/) also points to the existence of a certain limit number of sources. From these results, we can find the time at which the formation of quasars began. This was approximately one billion years after the Universe began expanding.

The exact nature of quasars is unknown at this stage and is very difficult to guess at. The discovery of the variable radio flux rendered all the conventional mechanisms of radio emission inadequate. The fairly rapid fluctuations of the radio flux are difficult to reconciliate with the emission mechanism of relativistic electrons moving in magnetic fields. Another alternative is to invoke coherent emission mechanisms (e.g., plasma oscillations /28/ or coherent stimulated emission of relativistic electrons /29/).

What is the probable structure of a quasar according to current notions? The core of a quasar is a nucleus measuring $\leq 10^{15}$ cm, whose mass is approximately 10^8 solar masses. The nucleus plays a definite role in the overall behavior of the quasar. In particular, its emission constitutes the main contribution to the continuous spectrum of the source. The nucleus is a giant star where the equilibrium is maintained by a balance between the gravitational energy and the energy of magnetic turbulent plasma or the rotational energy of the spinning star. The energy losses through the powerful radiation of the nucleus are made up by the gradual contraction of the star, i.e., by the gravitational energy resources. It follows from the theory of gravitational collapse that when a mass contracts to its gravitational radius $r_g = \frac{2CM}{c^2}$, it releases energy which amounts to several tens of percent of Mc^2 (thermonuclear reactions release only about 0.5% of Mc^2). For $M = 10^8 M_\odot$, $r_g = 3 \cdot 10^{13}$ cm, i.e., a figure of the order of magnitude of the diameter of the Earth's orbit. The energy resources corresponding to $(1/3)\,Mc^2$ are equal to $6 \cdot 10^{61}$ erg, which is sufficient to keep a quasar going for 20 million years at a rate of 10^{47} erg/sec. The existing estimates of quasar masses, however, are highly uncertain, and they probably provide only a lower limit (the mass of the nucleus is taken to be larger than the mass of the surrounding envelopes, which can be determined from emission and absorption lines). The activity of the nucleus is associated either with its pulsations or with the fact that it constitutes a close binary system of high-mass superstars. This activity involves ejection of ionized gas and streams of relativistic particles. It is quite probable that at the center of galaxies, and in particular at the center of our Galaxy, quasar-like objects exist. A certain region at the center of our Galaxy emits strong nonthermal radio emission.

The motion of ionized and neutral gas clouds in the central parts of galaxies is also reminiscent of quasars. A nucleus with bright emission lines was discovered at the center of the Andromeda Nebula (M 31). We have mentioned before the striking similarity in the optical and

radio spectra of the nucleus of the galaxy NGC 1275 and of quasars. However, effects of this kind in galaxies are many orders of magnitude less powerful than the corresponding phenomena in quasars.

It should be noted that the nature of many of the known radio sources is no less puzzling than the nature of quasars. For example, some double galaxies, including one of the brightest radio sources in the sky, Cygnus A, are great cosmic enigmas. The optical nebula located between the two radio sources is not a galaxy in the usual sense of the word. It seems to be made up entirely of high-temperature gas. Recently the radio galaxy has been shown to emit high amounts of energy in the X-ray spectrum. The radio galaxy Virgo A emits in the X-ray spectrum 100 times as much as in the radio and the optical spectrum /30/.

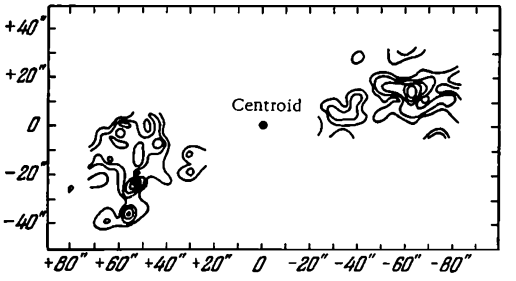


FIGURE 16. The structure of the radio galaxy Cygnus A at 11 cm.

Figure 16 is a chart of Cygnus A obtained at $\lambda = 11$ cm with a radio interferometer. The structure of this object is clearly very complex, and it contains several sources of small angular dimensions. Figure 17 is a photograph of the sky near the double radio source 3C 33 /31/. The radio source components show on the photograph as two ellipses which give an idea of the source size and correspond to a certain peripheral enhancement. Midway between the sources we see a galaxy, which apparently brought forth the two objects. Some of the puzzling questions are what caused this "ejection" from the galaxy, how to explain the striking likeness in the radio spectra of the ejected sources, what prevents this formation from expanding through the interstellar medium if these are indeed relativistic gas clouds, as many seem to think?

The most remarkable objects of this kind are the radio sources 3C 343 and 3C 343.1 /32/. Their spectra are also perfectly identical, the distance between the components is 29', the angular size of each component is less than 0".1. The parent galaxy has not been discovered so far. The identical spectra of two complex cosmic objects whose separation from one another is greater than the diameter of a sizeable galaxy are very difficult to account for by any of the known natural mechanisms.

Let us summarize. It is obviously too early to suggest that quasars or some of the radio galaxies are artificial sources of energy.

FIGURE 17. The area around the radio source 3C 33.

It seems, however, that this hypothesis definitely deserves more than a cursory glance. Anyway, this hypothesis has stimulated during recent years some discoveries of highly important properties of quasars, which are discussed in the next section. Observations have thus far established that quasars are the most powerful and yet the most compact energy sources among all known astrophysical objects (the quasar nucleus is smaller than the solar system whereas its radiation is more powerful than that of a thousand galaxies!). Future surveys in unmastered frequency ranges will show whether or not more powerful sources exist in the Universe. Studies of the most powerful objects will clearly enable us to fix an upper limit to the permissible energy output of a civilization.

Solid matter

From the point of view of modern physical concepts, the only state of aggregation of matter which is capable of storing indefinitely a large quantity of information is the solid state. The main feature of the solid state is the fixed and constant arrangement of atoms in the lattice. This feature is the basis of modern technology, in that it ensures constant and immutable properties of constructions; the same phenomenon made possible the development of biological processes on the Earth.

The solid matter also probably provides the basic constituent for the technology of supercivilizations, in particular in various data acquisition and processing systems. Therefore, a discovery of solid cosmic objects may have a significant bearing on the solution of our problem.

Unfortunately, the solid state of matter is the most difficult to detect in the Universe, because of its low temperature and the correspondingly weak emission of radiation. Therefore, our information on the quantity and properties of solid matter in the Universe is virtually nil.

The astrophysical data in our possession refer to planets and their satellites, to meteorites and interplanetary dust in the solar system (also measured from rockets), and to the extinction of stellar light by interstellar dust particles and its scattering in the reflecting nebulae. The properties of interstellar dust are mainly derived from theoretical considerations regarding the quantity of the heavy elements and the possible properties which cause mechanical destruction of the dust particles, their heating and cooling.

High-mass solid objects in the Universe are extremely intractable. Let us consider this point in some detail. There can be two different approaches to the search for these large solid objects: trying to detect the nearest individual massive objects and trying to detect the combined emission (or absorption) effect of a large assembly of solid bodies. Let d, δ, and T be respectively the size, the density, and the surface temperature of the solid objects, n the number of these objects in unit volume, l the size of that part of the Universe which is filled with these objects. The mean density of matter associated with the solid objects is then

$$\rho = n\,\delta d^3,$$

and the angular size of the nearest object is

$$\varphi_{max} = dn^{-1/3} = \left(\frac{\rho}{\delta}\right)^{1/3}, \tag{1.11}$$

the observed emission temperature of a large assembly of such objects (treated as the background emission) is

$$T_B = Tnd^2l, \tag{1.12}$$

and the optical thickness for light absorption or scattering by these solid objects is

$$\tau = nd^2l. \tag{1.13}$$

Assuming that the concentration of solid matter with $\delta \sim 1$ g/cm^3 does not exceed the mean density in the Universe $\rho \sim 10^{-29}$ g/cm^3 (for extragalactic solid objects) or the density in the Galaxy for galactic objects (either estimate is grossly exaggerated), we find that the angular size of the nearest objects does not exceed $4 \cdot 10^{-5}$ and $4 \cdot 10^{-3}$ sec, respectively. Since the surface temperature of these objects is limited, there is no way to detect them individually.

The combined emission of a large assembly of solid objects will be difficult to observe when $T_B \lesssim T_e$, where $T_e \sim 3$°K is the equilibrium temperature for all types of electromagnetic radiation in the Universe. Absorption effects are difficult to distinguish when $\tau \ll 1$. Since the surface temperature of the solid objects clearly should be greater than (or equal to) $T_e \approx 3$°K, the two conditions combined give the inequality

$nd^2l \ll 1$, and since $n = \frac{\rho}{\delta d^3}$, we find

$$d \gg \frac{\rho l}{\delta}; \tag{1.14}$$

assuming that the solid matter accounts for a noticeable fraction of the density in the Universe $\rho \sim 10^{-29}$ g/cm^3 (this estimate is a gross exaggeration), and taking for the density of the solid objects $\delta \sim 1$ g/cm^3 and $l \sim \frac{c}{H_0} \sim 10^{28}$ cm, we find that even in these extreme conditions particles measuring $d \gg 1$ mm remain absolutely undetectable.

In our opinion, this difficulty is virtually insurmountable in the sense that giant solid constructions of supercivilizations may remain undetectable even with the largest telescopes. The attempts to detect solid objects from their gravitation effects are also absolutely hopeless, since the existing estimates of the total mass of star clusters, galaxies, and clusters of galaxies are characterized by low accuracy (mainly because of the high proportion of low-luminosity stars). The estimates become more encouraging if we assume that the effective density of constructions in the Universe is $\delta \ll 1$. One of the examples of constructions of this kind is Dyson's sphere, a shell enclosing a star, with a radius of about 1 astronomical unit. The equivalent density of this construction is

$$\delta_{eq} \sim \delta \frac{4\pi r^2 \Delta}{\frac{4}{3}\pi r^3} = \frac{3\delta\Delta}{r},$$

where $r \sim 1.5 \cdot 10^{13}$ cm is the radius of the sphere, $\Delta \sim 10^2$ cm is the thickness of the sphere, and $\delta \sim 1$ g/cm^3. In this case, we find $\delta_{eq} \sim 2 \cdot 10^{-11}$ g/cm^3, the mass of the sphere is $M = \frac{4}{3}\pi r^3 \delta_{eq} \sim 3 \cdot 10^{29}$ g (approximately half the mass of the planet Saturn), and the visible angular dimensions (1.11) for the Metagalaxy and the Galaxy, respectively, are $\lesssim 0''.15$ and $\lesssim 15''$. These objects can be detected with modern telescopes. More detailed calculations /33, 34/ lead to the estimates listed in Table 1.2.

TABLE 1.2

P, W	10^{-11}			10^{-12}			10^{-14}		
D, cm	50	150	500	50	150	500	50	150	500
R, pc	2.06	6.46	20.6	6.58	20.1	65.8	65.8	102	658

In the calculations of Table 1.2 it was assumed that the heat emission of the sphere (with a surface temperature of 300°K) at wavelengths between 8 and 13 microns was observed with modern high-sensitivity bolometers and optical telescopes. In this table, P is the bolometer sensitivity, D is the telescope diameter, and R is the maximum distance at which the thermal emission of Dyson's sphere is detectable, assuming a minimum signal/noise ratio of 9. Note that these observations can be easily carried

out with the existing telescopes on Earth, since the 8—13μ range corresponds to one of the transparency windows of the Earth's atmosphere. In general, the thermal emission peak of solid objects at temperatures between 3 and 300°K falls between 10μ and 1 mm, and at these wavelengths the atmosphere is highly opaque, mainly due to the absorption by water vapor. The search for these sources should therefore lean heavily on observations from beyond the atmosphere.

§6. THE SEARCH FOR INFORMATION TRANSMISSIONS

In the previous section we discussed the search for the various signs of activity of civilizations. One of the most probable elements of this activity is apparently transmission and exchange of information. These transmissions can be divided into two broad types: 1) exchange of information between highly developed civilizations of approximately the same level, and 2) transmission of information aimed at raising the level of less developed civilizations. If supercivilizations actually exist, transmissions of the first group may prove virtually inaccessible to us (e.g., these transmissions may be directed by tight-beam systems and the transmission line need not necessarily intercept the solar system). On the other hand, transmissions of the second group, by their very nature, should be readily accessible and easily detectable by others. The reception of transmissions of this kind is expected to have a fundamentally significant influence on the development of our civilization (von Hoerner's feedback effect /2/), and as a result we will rapidly rise to the highest level of civilization currently existing in the Universe. Probably the fastest and the simplest (though the most fantastically sounding) way to achieve this advancement is by merging with the nearest supercivilization.

How is the search for transmissions of this kind to be planned?

Redundancy considerations indicate that we can hardly expect a great number of transmissions of this kind. We will do better to concentrate on one or several sources of electromagnetic radiation which stand out among the rest in terms of their intensity or some other property. The search for these prominent sources can be effected by means of sky surveys in the least noisy frequency range. As we have noted above, the technical means for the detection of electromagnetic radiation have improved to such an extent that instrumental noise is no longer the main limiting factor. In the next 5— 10 years, apparently, the receivers will attain their maximum sensitivity for astrophysical work, which is determined in each frequency range by the intensity of background radiation and by random fluctuations of the signal. Figure 18 plots the background intensity spectrum for an observer situated in the intergalactic space, far from the bright galaxies. This spectrum has been reconstructed from the results of measurements in the radio, optical, and partly X-ray spectra, and also between the optical and the X-ray spectra the curve is based on theoretical calculations, which take into consideration the emission of the interstellar dust in galaxies and the total emission of the galactic stars, and also on extrapolation of the available results

of observations /35, 36/. For an Earth-bound observer, the background radiation of our Galaxy has to be added to this spectrum. The resultant background intensity from the Earth is shown in Figure 19 for the "brightest" (the center of the Galaxy) and the "coldest" (the Galactic pole) parts of the sky.

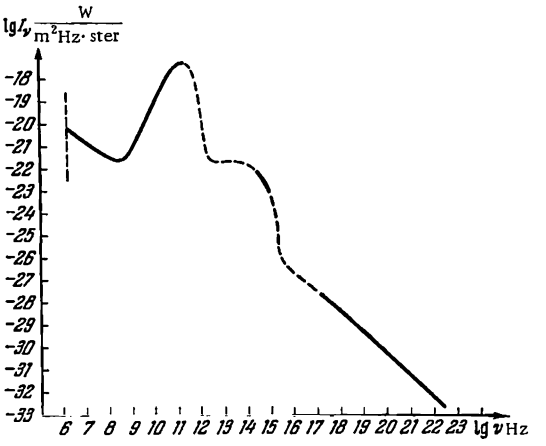


FIGURE 18. The spectrum of the background electromagnetic radiation for an observer in the intergalactic space.

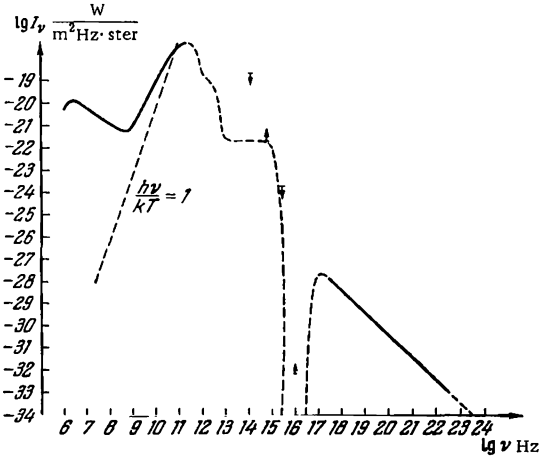


FIGURE 19. The spectrum of the background electromagnetic radiation for an observer in the solar system.

Both spectra show deep intensity minima, and these valleys are apparently the most suitable for interstellar communication. The discrete (quantized) nature of the electromagnetic radiation is another significant

factor to be considered in connection with the choice of the transmission range. The distinctive feature of the spectra in Figures 18 and 19 is that the background intensity is everywhere higher than the blackbody intensity at 3°K. The range with the minimum equivalent blackbody temperature (the region where the "relic" background radiation predominates) lies at wavelengths between 3 m and 30 cm in Figure 18 (this range is somewhat narrower for the case in Figure 19). In both figures, the dashed line marks the limit where for a blackbody radiation $\frac{h\nu}{kT} = 1$, and consequently $I_\nu = B_\nu = \frac{2h\nu^3}{c^2}\frac{1}{e^{\frac{h\nu}{kT}}-1} = \frac{2h\nu^2}{c^2}\frac{1}{e-1}$. To the right of this limit we have $\frac{h\nu}{kT} > 1$ and the quantum effects are therefore most prominent.

The solution of the problem of optimum signal transmission against a noisy background essentially depends on the particular parameters that are to be optimized. Allowance for quantum and classical fluctuations leads to the following expression for the maximum quantity of information which can be received in unit time in a unit frequency interval /37/:

$$c_\nu = \ln\left[1+\frac{P_\nu}{h\nu}\left(1-e^{-\frac{h\nu}{kT}}\right)\right] + \left[\frac{P_\nu}{h\nu}+\frac{1}{e^{\frac{h\nu}{kT}}-1}\right]\ln\left[1+\frac{1}{\frac{P_\nu}{h\nu}+\frac{1}{e^{\frac{h\nu}{kT}}-1}}\right] - \frac{\frac{h\nu}{kT}}{e^{\frac{h\nu}{kT}}-1}. \quad (1.15)$$

Here P_ν is the power spectrum of the received signal at the receiver input, $T = T(\nu)$ is the effective temperature of all the noises corresponding to an approximate input noise spectrum of the form $\varepsilon_\nu = \frac{h\nu}{e^{\frac{h\nu}{kT}}-1}$. For fixed P_ν and T, the function c_ν decreases monotonically with increasing frequency (Figure 20).

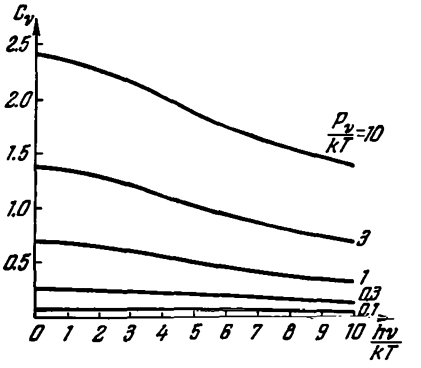


FIGURE 20. C_ν vs. frequency.

In the classical case $\left(\frac{h\nu}{kT} \ll 1\right)$, equation (1.15) takes the form

$$c_\nu = \ln\left(1+\frac{P_\nu}{kT}\right) - \frac{1}{24}\left(\frac{h\nu}{kT}\right)^2\left(\frac{P_\nu}{P_\nu+kT}\right)^2 + \ldots \quad (1.16)$$

If we retain only the first term of the expansion, (1.16) coincides with Shannon's standard expression for the rate of information transmission.*

In another limiting case $\frac{h\nu}{kT} \gg 1$, and assuming that the signal is much more powerful than the noise $\left(P_\nu \gg \frac{h\nu}{e^{\frac{h\nu}{kT}}-1}\right)$, we get

$$c_\nu = \ln\left(1+\frac{P_\nu}{h\nu}\right) + \frac{P_\nu}{h\nu}\ln\left(1+\frac{h\nu}{P_\nu}\right), \tag{1.17}$$

i.e., the rate of information transmission depends only on the number of quanta of the signal received in unit time $\frac{P_\nu}{h\nu}$. For high spectral intensities of the signal $\left(\frac{P_\nu}{h\nu} \gg 1\right)$, equation (1.17) takes the form

$$c_\nu = \ln\left(1+\frac{P_\nu}{h\nu}\right) + 1, \tag{1.18}$$

which is also close to Shannon's expression if we introduce the "equivalent temperature" of the quantum noise $kT_{eq} = h\nu$.

For an ideal detector, the noise ε_ν is determined by the intensity of the sky background radiation I_ν, i.e.,

$$\varepsilon_\nu = \frac{1}{2} I_\nu \Omega_\nu A_\nu = \frac{c^2 I_\nu}{2\nu^2}. \tag{1.19}$$

Here A_ν and Ω_ν is the effective collecting surface and the effective solid angle of the receiver at frequency ν (the antenna, if radio frequencies are considered), and the factor $1/2$ allows for one polarization component of the intensity I_ν (both components are assumed to have the same intensity). The last part of equality (1.19) is valid only if $A_\nu\Omega_\nu = \lambda^2$, which is not always true. For example, for the optical telescopes, the minimum solid angle Ω_ν is generally determined by the scattering in atmospheric inhomogeneities, and not by the diffraction pattern of the point source. As a result, the angular size of the source is very seldom less than 1", and the adjustable aperture used to restrict the background should not be less than this figure. Thus, we always have $A_\nu\Omega_\nu \geqslant \lambda^2$ and the equality in this relation (corresponding to a pure diffraction image) ensures the best signal/noise ratio.

The signal power spectrum at the detector input is related to the radiation flux of a point source F_ν by the equality

$$P_\nu = \frac{1}{2} A_\nu F_\nu, \tag{1.20}$$

where the factor $1/2$ allows for the polarization components, as in (1.19).

These general relations make it possible to assess the peculiar features of signals of artificial origin.

The reception of signals from extraterrestrial civilizations can be divided into three stages: 1) search for call signals and their decoding, 2) search for the key to transmission and its decoding, 3) reception and decoding of information.

Let us consider in some detail the first phase of the procedure, namely the search for call signals and the choice of the most suitable frequency range.

* See also Chapter III.

Call signals are intended to facilitate the detection of the source, and they carry a certain minimum quantity of information which is sufficient to firmly identify the source as an artificial object.

The choice of the optimum frequency range for call signals thus amounts to the following. We have to find the frequency ν and the operating conditions of the receiver which ensure the maximum signal/noise ratio for a given total energy flux from the source per unit surface area near the Earth F and the given search time t_0.

If an ideal receiver is used, the root-mean-square noise power at the input is determined by the fluctuations associated with the natural background radiation from outer space. Allowing for the fluctuation in the number of photons, we have

$$\sqrt{\overline{\Delta P_\nu^2}} = [\varepsilon_\nu^2 + \varepsilon_\nu h\nu]^{1/2} \left(\frac{\Delta\nu}{\tau}\right)^{1/2}, \tag{1.21}$$

where ε_ν is defined by (1.19), and $\Delta\nu$ and τ are the receiver band width and time constant, respectively. Seeing that the input signal power is $P_s = P_\nu \Delta\nu = \frac{1}{2} F A_\nu$, we find for the signal/noise ratio

$$N = \frac{\frac{1}{2} F A_\nu \sqrt{\tau}}{[\varepsilon_\nu^2 + \varepsilon_\nu h\nu]^{1/2} \sqrt{\Delta\nu}}. \tag{1.22}$$

The entire search time t_0 incorporates both the frequency search and the direction search, so that we have to maximize N for a given t_0,

$$t_0 = \tau \frac{\nu}{\Delta\nu} \frac{4\pi}{\Omega_\nu}. \tag{1.23}$$

As we have noted before, the real reception conditions are such that the solid angle Ω_ν of the receiver and the effective collecting area A_ν are related by the equality

$$\Omega_\nu A_\nu = k_\nu \lambda^2, \tag{1.24}$$

where the numerical coefficient $k_\nu \geqslant 1$. Because of the great difficulties in the manufacture of large precision surfaces, the coefficient k_ν increases with the increase in frequency. The actual conditions of propagation of light and radio waves in the atmosphere also increase the coefficient k_ν, and this effect is particularly pronounced for observations in the optical region. In observations from outside the atmosphere, allowance should be made for the increase of angular dimensions due to the propagation of radio waves in the interstellar and the intergalactic plasma. This problem is discussed in detail in Chapter II, where it is shown that the scattering is negligible in the centimeter and the decimeter range, but it may reach significant values for the meter wavelengths.

Using (1.19), (1.23), and (1.24), we write (1.22) in the form

$$N = \frac{F\sqrt{A_\nu \nu t_0}}{\sqrt{4\pi}\, c \sqrt{I_\nu^2 k_\nu + I_\nu \frac{2h\nu^3}{c^2}}}. \tag{1.25}$$

The following conclusions can be drawn from this relation. First, to obtain the maximum N, we should make k_ν as small as possible, if the main contribution comes from classical fluctuations $\left(I_\nu k_\nu \gg \frac{2h\nu^3}{c^2}\right)$. This condition is always imposed in the region where the background is described by an effective temperature which satisfies the inequality $\frac{h\nu}{kT} \ll 1$. However, in the quantum region $\left(\frac{h\nu}{kT} \gg 1\right)$, in the case of large k_ν, the first term in the radicand in the denominator of (1.25) may become significant. Moreover, for $I_\nu k_\nu \ll \frac{2h\nu^3}{c^2}$, which applies only to the short wave region (where $\frac{h\nu}{kT} \gg 1$), N is independent of k_ν. Finally, the requirement of maximum N from (1.25) does not impose any requirements on the band width $\Delta\nu$ and the time constant τ.

For the region where the background intensity is described by the classical Rayleigh–Jeans formula $\left(\frac{h\nu}{kT} \ll 1\right)$ we have from (1.25)

$$N_c = \frac{F\sqrt{A_\nu \nu I_0}}{\sqrt{4\pi}\, c I_\nu \sqrt{k_\nu}}, \tag{1.26}$$

and in the quantum region $\left(\frac{h\nu}{kT} \gg 1 \text{ and } I_\nu k_\nu < \frac{2h\nu^3}{c^2}\right)$ we have

$$N_q = \frac{F\sqrt{A_\nu \nu I_0}}{\sqrt{4\pi}\, c \sqrt{I_\nu \frac{2h\nu^3}{c^2}}}. \tag{1.27}$$

Let us again return to Figures 18 and 19, where the dashed line marks the limit of the two regions for $k_\nu \sim 1$. In the general case, the first and second term in the denominator of (1.25) are equal for $I_\nu = \frac{2h\nu^3}{c^2 k_\nu}$. If $k_\nu \gg 1$, the boundary between the quantum and the classical region is markedly shifted in the short-wave direction.

Quantitative estimates based on (1.25)–(1.27) lead to a definite conclusion regarding the optimum frequency range: the decimeter range of wavelengths, where the radio background is minimum ($\lambda \sim 10-50$ cm), ensures the maximum signal/noise ratio for sky surveys during a given time τ.*

In addition to being easy to detect, call signals should contain a minimum quantity of information which will label them as artificial signals. The fundamental differences between signals of natural and artificial origin have not been defined yet. These differences, however, are reflected mainly in the information content of the signals, and not in their shape. Transmissions, and even call signals, should carry certain information which is absent in the radiation generated by natural processes.

Another question to ask is, shall we be able to understand the communications received from civilizations whose age and evolution are

* The latest observations of a new type of object – pulsars – indicate that there exists still another type of noise in ultra-long-range transmissions. This noise, attributed to fluctuations of the refractive index of the interstellar plasma, makes the signal disappear for long stretches of time. This effect has been poorly studied at this stage. Unlike the background radiation, this is a multiplicative noise, and it will probably shift the optimum frequency range toward shorter wavelengths.

substantially different from those of our civilization? There is clearly room for understanding if a single common language can be devised. The uniform structure of the Universe and the universality of the laws of nature in different places and different times, as they emerge from observational data, seem to provide this common language.

We are now in a position to summarize our conclusions regarding call signals. Empirical considerations show that the quantity of information needed to label a signal as artificial should contain more than 10 and less than 100 bits:

$$10 < J < 10^2. \tag{1.28}$$

Since the laws of nature are universal, the best policy would be to transmit a certain combination of digits as a call signal of minimum information content. For example, only 60 bits are required to transmit in hexadecimal binary code the first eight primary numbers, their sum, and the space signal between successive transmissions: 000001, 000010, 000011, 000101, 000111, 001011, 001101, 010001, 111011, 000000, ... which stands for 1, 2, 3, 5, 7, 11, 13, 17, 59, 0, ... A periodically repeated transmission of this kind will leave no doubt whatsoever regarding its artificial origin.

There is a great variety of different call signals. Measurements of electromagnetic radiation record the following parameters: the two spatial coordinates of the source, the time of observation, the frequency, the intensity, the degree of linear polarization and its position angle, the degree of circular polarization and its position angle. In principle, a change in any of these parameters as a function of a change in any other parameter may be regarded as a source of information. The different call signals are conveniently divided into two groups: transient call signals and stationary (or steady-state) call signals. Transient call signals involve a time variation in any of the above parameters (e.g., the binary code can be transmitted by altering the sense of circular polarization). Stationary call signals involve a regular variation of one parameter as a function of another, irrespective of the time factor. For example, the variation of the sense of circular polarization as a function of frequency may contain the minimum quantity of information (1.28).

It is not clear at present which of the different transmission techniques is the most effective. Therefore, no exact criteria are available for analyzing the parameters of suspicious sources.

Let us consider still another possibility of searching for call signals. In all likelihood, only a minor fraction of the transmitter power is used up in sending special call signals. Is it not possible to use certain general properties of the transmitted information as a built-in call signal? If the transmission covers a very wide frequency band, the averaging effect may increase the measurement sensitivity several orders of magnitude compared to the sensitivity of narrow-band measurements without averaging. Thus, in radiometric measurements of the mean source power, the signal/noise ratio increases by a factor of $n = \sqrt{\Delta\nu\tau}$ compared to its value in measurements without averaging. Therefore, to search for a source transmitting in a band $\Delta\nu$, we need an antenna with $1/n$ the effective area needed for receiving information

from the same source. As an example, if the information is transmitted in a band of 10^{11} Hz and the averaging time is 10 sec, we have $n = 10^6$.

Let us now consider the problem of optimal transmission of information. The main question is how the transmitter energy should be distributed over the spectrum to ensure the maximum transmission rate. The fixed factors to be considered are the noise intensity spectrum I_ν and the total energy flux from the transmitter per unit surface area at the Earth, F. Problems of this type /37/ are solved by varying relation (1.15) under the fixed conditions. The optimum source spectrum is found to be

$$F_\nu = \frac{\frac{2h\nu}{A_\nu}}{e^{a\frac{2h\nu}{A_\nu}} - 1} - I_\nu \Omega_\nu, \tag{1.29}$$

where a is determined from the condition $\int_L F_\nu d\nu = F$. Smaller values of a correspond to larger values of F. Seeing that by (1.24) $\Omega_\nu A_\nu \geqslant \lambda^2$, we find

$$F_\nu \leqslant \frac{2h\nu}{A_\nu}\left[\frac{1}{e^{a\frac{2h\nu}{A_\nu}} - 1} - \frac{1}{e^{\frac{h\nu}{kT}} - 1}\right], \tag{1.30}$$

and since $F_\nu \geqslant 0$, we have

$$\frac{A_\nu}{2ak} > T, \tag{1.31}$$

where T, as before, is the effective background temperature of frequency ν. We thus reach the following conclusion: the optimum transmission range corresponds to that part of the spectrum where the effective background temperature is minimum.*

For the background electromagnetic radiation from outer space, this region corresponds to the frequencies where the so-called relic radiation prevails, i.e., the radiation described by Planck's formula with $T \sim 3°K$. This range covers the spectrum from submillimeter to decimeter wavelengths, with a background intensity maximum near $\lambda \sim 1.7$ mm (see Figures 18 and 19).

A more definite shape of the source spectrum can be derived using the dependence of A_ν and Ω_ν on frequency. Let us consider two possible cases.

1. $A_\nu = A\nu^{-2}$, $\Omega_\nu = \Omega = \text{const}$. As we have noted in § 3, this case corresponds to the limitations imposed on the largest possible antennas, provided that the relative surface finishing accuracy is approximately the same at all wavelengths. The shape of the spectrum F_ν depends on the parameter F (Figure 21). For small F (i.e., low-power transmitters), the maximum F_ν corresponds to the minimum background intensity in the decimeter range, i.e., it coincides with the best frequencies for call signals. For high F, the transmitter spectrum is broader, and, if the background radiation is negligible, we have

$$F_\nu = \frac{\frac{2h\nu^3}{A}}{e^{a\frac{2h\nu^3}{A}} - 1}. \tag{1.32}$$

* The entire range, however, may shift toward shorter wavelengths due to the factors mentioned in the footnote on p.47.

This spectrum is characterized by a plateau in the low-frequency region and an exponentially falling branch (no maximum) at high frequencies. The shoulder is associated with information losses due to quantum fluctuations of the signal.

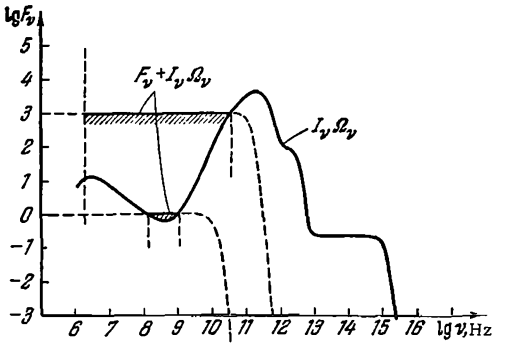


FIGURE 21. Energy distribution in the spectrum of an artificial source for $A_\nu \propto \nu^{-2}$, $\Omega_\nu = \text{const}$.

2. $A_\nu = A = \text{const}$, $\Omega_\nu = \frac{c^2\nu^{-2}}{A}$. This case corresponds to measurements with a single antenna, which receives the entire spectrum of the signal:

$$F_\nu = \frac{2h\nu}{A}\left(\frac{1}{e^{\frac{2h\nu a}{A}} - 1} - \frac{1}{e^{\frac{h\nu}{kT}} - 1}\right). \tag{1.33}$$

For small F, we may write $\frac{A}{2h\nu a} = \frac{k(T+\Delta T)}{h\nu}$, $\Delta T \ll T$. Then

$$F_\nu = \frac{2k\,\Delta T}{A}\left(\frac{h\nu}{kT}\right)^2 \frac{e^{\frac{h\nu}{kT}}}{\left(e^{\frac{h\nu}{kT}} - 1\right)^2}. \tag{1.34}$$

For $\frac{h\nu}{kT} \ll 1$, the spectrum has a plateau, $F_\nu = \frac{2k\,\Delta T}{A}$; then the flux increases, reaching a maximum at $\frac{h\nu}{kT} \sim 1.1$ (for T = 3°K, this corresponds to λ = 4.8 mm), and then falls of exponentially. The maximum F_ν is a factor of 2.7 greater than the plateau value.

As F increases, the background limitations become progressively less significant, and the spectrum width increases. At very high transmitter powers, the distribution shows a plateau on the low-frequency side and falls off exponentially (no maximum) at high frequencies:

$$F_\nu = \frac{2h\nu}{A}\frac{1}{e^{\frac{2h\nu a}{A}} - 1}. \tag{1.35}$$

The expected spectrum curve (qualitative picture) for various F is given in Figure 22.

Let us now consider the general properties of information transmission in order to find some built-in criteria for a preliminary selection of radio sources and a search for call signals.

1. A significant part of the spectrum of an artificial source invariably falls in the radio frequency range, with a maximum at the frequencies corresponding to the minimum background intensity (the short-wave part of the decimeter range) or in the millimeter range. A spectrum with a maximum at decimeter wavelengths or with a plateau in this range and a maximum at millimeter wavelengths provides a tentative criterion for the selection of suspicious objects.

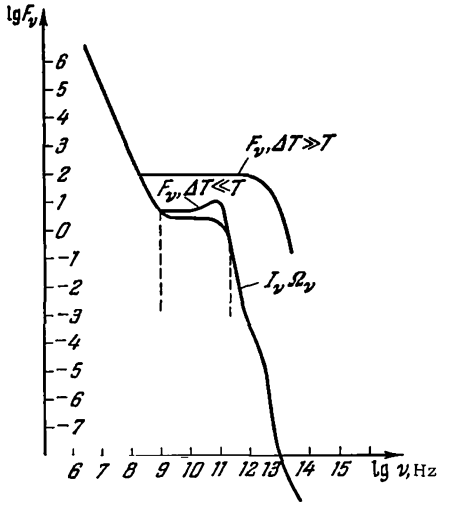


FIGURE 22. Energy distribution in the spectrum of an artificial source for $A_\nu = \text{const}$, $\Omega_\nu \sim \nu^{-2}$.

2. Minimum angular size of the suspects (radio sources) may also be regarded as a very strong tentative criterion.

3. Measurements of other astrophysical parameters of the source in other spectral regions can also be used for preliminary selection (circular polarization, optical and radio lines, optical identification, X-ray emission, etc.).

In this respect, the search for artificial sources is virtually coincidental with the general trend of modern observational radio astronomy. It is probably for this reason that the discussion of the possible tentative criteria for the identification of artificial sources left a profound imprint on radio astronomical work. Thus, during the 1964 discussions surrounding the program of search for extraterrestrial civilizations /38/ it was first suggested that artificial sources should have a spectrum with a maximum at decimeter and centimeter wavelengths, minimum angular size, and definite variability with time. CTA-102 was mentioned as a probable suspect meeting these criteria. In the years that followed, the relevant properties were discovered for a number of sources, CTA-102 included. New sources with radio emission concentrated mainly in the decimeter range were discovered. One of the most remarkable objects in this respect is the source 1934 − 63 (coordinates $\alpha = 19^h34^m48^s.9$, $\delta = -63°49'42''$ (1950)) /39/. Figure 23 shows the spectrum of this object, with a maximum around $\lambda = 21$ cm. Figure 24 is a photograph of the sky area showing this source. A galaxy with a bright star-like nucleus is observed at the same position, and it is joined by a hardly visible bridge to another star-like object.

A striking example of a source with a flat plateau spectrum and a probable maximum in the millimeter or submillimeter range is 3C 273B (see Figure 14). This and a number of other sources have extremely small angular size (less than 0".002) and their radio flux is highly variable.

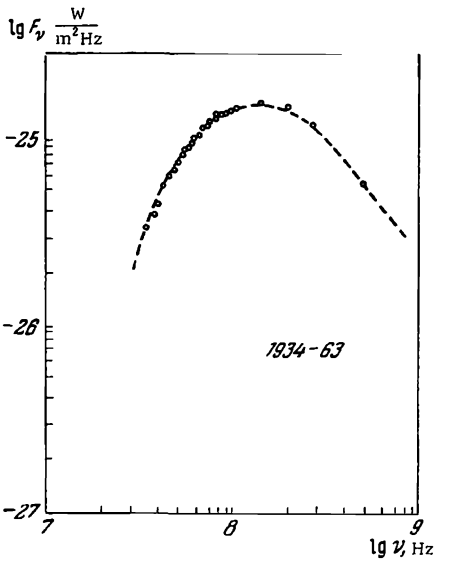


FIGURE 23. The spectrum of the radio source 1934–63.

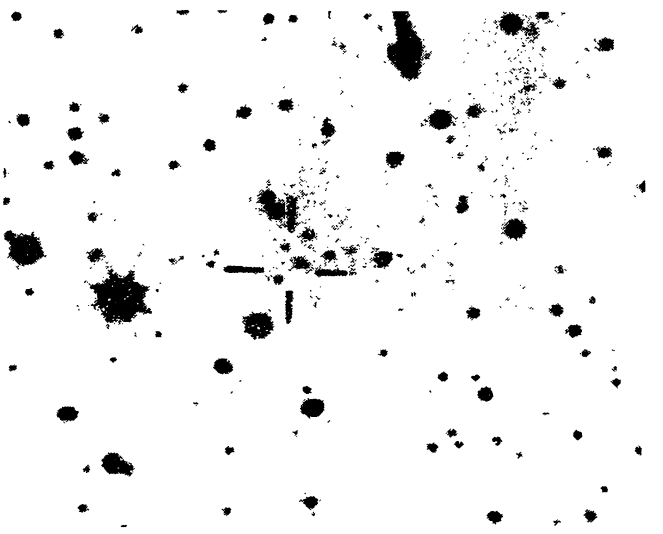


FIGURE 24. The sky area around the radio source 1934–63.

As we have already noted, these observational results are inconsistent with the synchrotron radiation mechanism generally used for radio sources. Calculations show that this mechanism will fail to generate the observed power in sources of such small size. Therefore, processes associated with collective coherent emission (plasma oscillations /28/, stimulated

emission /29/) are currently invoked to explain the observed effects. In this connection, note that the radiation from an artificial transmitter is a typical example of coherent emission.

The discovery of the anomalous strong line radiation at wavelengths near 18 cm is another important factor which has bearing on our problem /40/. The lines at 1612, 1665, 1667, and 1720 MHz are the splitting components (Λ-doubling and hyperfine structure of the lowest energy level of the hydroxyl molecule OH. Observations reveal the existence of an unusually powerful radiation in these lines (especially at 1665 and 1667 MHz) from regions of very small angular dimensions inside ionized gas clouds. Figure 25 is a photograph of one of these nebulae (NGC 6334); the squares mark the regions of anomalously strong monochromatic radiation /43/.

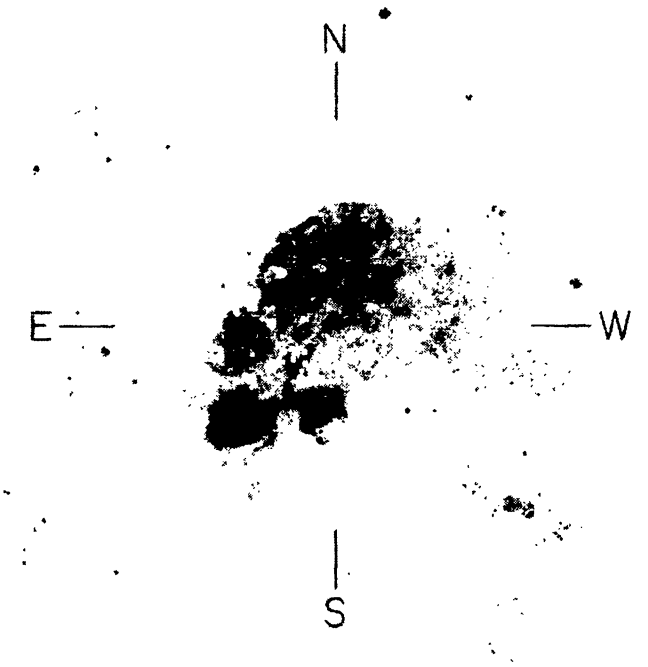


FIGURE 25. Nebula NGC 6334 showing regions of OH line emission.

The very existence of OH molecules in H II regions, where the temperature is around 10,000°, is in itself a highly surprising fact. Moreover, this emission has quite unusual properties. The angular size of the emitting regions is less than 0".002 (linear size less than 4 a.u.). We can therefore only give an upper bound estimate of the effective temperature at the peak of the line profile, which turns out to be over 10^{13} degrees. At the same time, the unusually narrow line profile (less than 400 Hz in some cases) points to a temperature not exceeding 10°K. This relationship between intensity and line width is possible only in nonlinear emission processes, not

unlike the generation mechanism of molecular masers and lasers in the laboratory. Further measurements of the interstellar hydroxyl lines revealed an almost 100% circular polarization of the strongest components; in some cases, strong linear polarization is also observed. Some of the lines show pronounced variation of the component intensities from day to day. Figure 26 is the profile of the 1665 MHz line of the nebula W 49 in linearly polarized, right-hand polarized, and left-hand polarized radiation /41/. Figure 27 shows the profile of the same line of the nebula NGC 6334 on different days /42/.

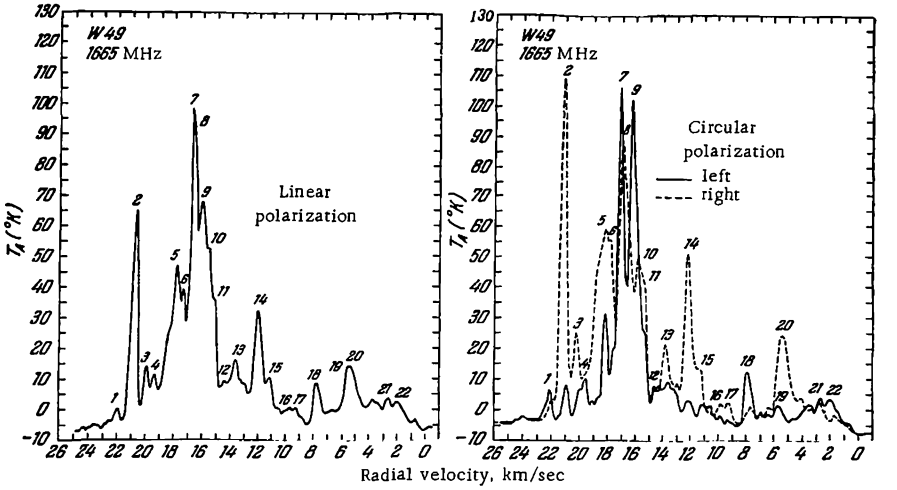


FIGURE 26. The 1665 MHz line profile of W 49 for linearly polarized, right-hand polarized, and left-hand polarized radiation.

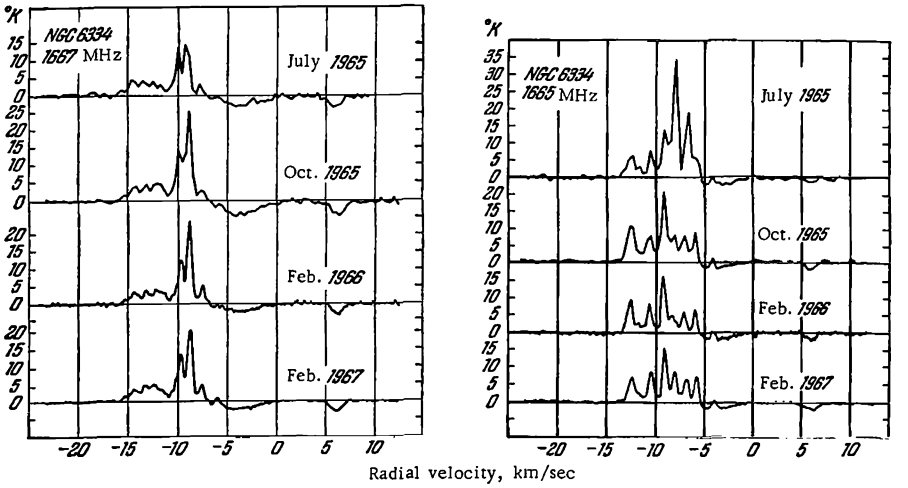


FIGURE 27. The 1665 MHz line profile of NGC 6334 according to observations on different days.

Theoretical estimates of the possibilities of population inversion of the energy levels in atoms and molecules corresponding to radio-frequency transitions show this to be a most likely event under the natural conditions in the interstellar medium /43/. Nevertheless, the exact mechanism of this stimulated emission is not clear today.

It is, however, important to remember in our search for call signals that the natural conditions in the interstellar space may greatly simplify in this respect the problem of creation of ultra-powerful narrow-band radio generators.

§7. THE PROGRAM OF SEARCH FOR SUPERCIVILIZATIONS

In the preceding sections we tried to justify our thesis according to which only the search for signals and signs of activity of supercivilizations can be carried out with useful results in the next few years. Over a period of some 10 years, astronomers can collect enough information about all the brightest sources in all the regions of the electromagnetic spectrum. This task is coincidental with the main trend of astrophysical research today. However, now is the time to propose some sort of a specific program for a search for artificial sources. There are reasons to believe that transmission of information is one of the basic conditions of existence for supercivilizations. We should therefore develop a special program for the detection of call signals accompanying these transmissions. Our analysis shows that the most likely frequency range for the call signals can be identified with fair certainty. The other parameters (frequency band, length of transmission, polarization, etc.), however, are very difficult to guess at beforehand. If the transmitted quantity of information is very large, we should naturally expect a very wide band transmission, so that we will have to look for artificial sources among a multitude of natural radio sources. It is quite probable, however, that the low-intensity wide-band transmission carrying the bulk of information is accompanied by powerful call signals generated at fixed frequencies in a very narrow band or in the form of very brief and yet powerful pulses.

Let us list the main directions of research which are of the greatest interest from the point of view of the search for call signals:

1. Sky surveys at 3, 10, 30, 100, and 300 microns and especially at 1, 3, 10 mm and 3 and 10 cm aimed at discovering at least 100 of the brightest sources in each frequency range.
2. Detailed studies of the properties of quasars and other "suspicious" objects.
3. Search for anomalously powerful monochromatic radio sources (like the hydroxyl line emission) in the decimeter range.
4. Search for pulse signals of interstellar origin in the same range.*
5. Search for monochromatic signals of variable frequency in the same range.

* The search for these signals began in 1967, with the discovery of pulsars. Among the known sources of this type, however, there are still no indications of artificiality.

The currently available information on each of the five items above is pitiful and negligible compared to what could have been collected with a properly planned utilization of modern means. The preliminary criteria discussed in the previous section may prove to be of great help in the preliminary sifting through of "suspects." The angular size criterion is particularly useful. The angular dimensions can be accurately measured with a radio interferometer. Radio interferometers with a base of the order of the Earth's diameter are currently available for the centimeter and decimeter wavelengths /44/. In the near future, one of the antennas will probably be mounted on an interplanetary spacecraft, thus giving a radio interferometer with a base comparable with the dimensions of the Earth's orbit. Other promising directions include the estimates of the maximum linear dimensions from the time variation of one of the source parameters (e.g., radiation flux or polarization). Since the velocity of light is finite, the radiation of the entire object can be observed to change simultaneously in a time t only if $t > \frac{r}{c}$, where r is the radius of the object. Suppose that the quasar 3C 273B is an object of mass $M \sim 10^8$ solar masses and its radius is greater than the critical (gravitational) radius $r_g = \frac{2GM}{c^2} = 3 \cdot 10^{13}$ cm (G is the gravitational constant). We thus come to the conclusion that the brightness of this source cannot change faster than with a period of $T \sim \frac{r_g}{c} = 10^3$ sec. Any discovery of faster light variation would point to a smaller mass and radius of this object.

Preliminary selection using the tentative criteria is a necessary, though not sufficient, stage of the general search procedure. Once a sufficient number of "suspects" have been selected, we have to start looking for "meaningful contents" in the radiation from these objects. This work, supported by parallel theoretical analysis of the various alternatives, will help to improve the future search program. In particular, we hope that significant information on the parameters of quasars and their time variation in various spectral regions will be accumulated in the course of the international program launched in 1966 /45/.

At present, we have no theory to enable us to assess the presence or the absence of meaningful information in the received signals. Man is the only suitable candidate for making decisions in this direction, and we are thus inevitably faced with the difficulties of subjective approach to the search program. This approach, however, will not be entirely arbitrary. A certain measure of objectivity will be derived from the observed universality of the laws of nature and their constancy in space and time. The universal laws of nature can be used as a common basis of understanding with other civilizations and, in particular, enable us to develop an objective search program. In principle, we can probably devise a procedure and build an analyzing machine for the comparison of the known universal laws of nature (mathematical relations in the simplest case) with any information received from outer space. In our opinion, this problem is definitely solvable, at least as far as the search for call signals is concerned.

Bibliography

1. Shklovskii, I. S. Vselennaya, zhizn', razum (Life and Intelligence in the Universe) 2nd Ed. — "Nauka." 1965.
2. Cameron, A. (Editor). Interstellar Communication. — New York. Benjamin. 1963.
3. Shklovskii, I. S. and C. Sagan. Intelligent Life in the Universe. — Holden Day. 1966.
4. Vologdin, A. G. Pervye shagi evolyutsii (The First Steps of Evolution). — Literaturnaya gazeta, 1 February, 1967.
5. Baranov, V. I. — Astron. Zhurnal, Vol. 43:1074. 1966; Fisher, D. E. — 20th IUPAC Congress, Moscow. 1965.
6. Gerling, E. K., V. A. Maslennikov and I. M. Morozova. — Ibid.
7. Nablyudatel'nye osnovy kosmologii (Observational Principles of Cosmology). Collection of articles. — Mir. 1965.
8. Burbidge, E. M. Quasi-stellar Objects. — San Diego, Univ. of California, California. 1967.
9. Lyapunov, A. A. — Problemy Kibernetiki, No. 10:179. 1963.
10. Kardashev, N. S. and G. B. Sholomitskii. — Astr. Tsirk., No. 336. 1965.
11. Rose, D. and M. Clark. Plasmas and Controlled Fusion. — Cambridge, Mass. MIT Press. 1961.
12. Andrillat, Y. and M. Andrillat. — Publ. de l'Observatoire de Haut Provence, Vol. 7, No. 11. 1964.
13. Dibai, E. A. and V. I. Pronik. — Astr. Tsirk., No. 286. 1964.
14. Ozernoi, L. M. and V. E. Chertoprud. — Astron. Zhurnal, Vol. 43:20. 1966.
15. Geyer, E. — Zs. für Astroph., Vol. 60:112–114. 1964.
16. Kinman, T., E. Lamla, and C. Wirtanen. — Contr. from Lick Observatory, No. 225. 1966.
17. Sandage, A., J. Westphal, and P. Srittmatter. — Ap. J., Vol. 146:332. 1966.
18. Hoerner, S. von, — Ap. J., Vol. 144:483. 1966.
19. Adgie R., H. Gent, O. Slee, A. Frost, H. Palmer, and B. Rowson. — Nature, Vol. 208:275. 1965.
20. Cohen, M., E. Gunderman, M. Hardebeck, and L. Sharol. — Sky and Telescope, Vol. 34: 143. 1967.
21. Kellerman, K. and I. Pauling — Toth. — Ap. J., Vol. 146. 1966.
22. Berge, G. and G. Seielstad. — Observations of the Owens Valley Radio Observatory, No. 9. 1966.
23. Low, F. — Ap. J., Vol. 142:1287. 1965; Ap. J., Vol. 150. 1967.
24. Friedman, H. and E. Byram. — 7 Cospar Symposium, London, 24–28 July. 1967.
25. Shklovskii, I. S. — Astron. Zhurnal, Vol. 42:893. 1965.
26. Sandage, A. and W. Muller. — Ap. J., Vol. 144:1240. 1966.
27. Veron, P. — Ann. d'Astrophys., Vol. 29:231. 1966.
28. Ginzburg, V. L. and M. M. Ozernoi. — Ap. J., Vol. 144:599. 1966.
29. Zheleznyakov, V. V. — ZhETF, Vol. 57:570. 1966; Astron. Zhurnal, Vol. 44. 1967; Kaplan, S. A. — Astrofizika, Vol. 2:409. 1966.
30. Byram, E., T. Chabb, and H. Friedman. — Science, Vol. 152:66. 1966.
31. Moffet, A. — Annual Review of Astronomy and Astrophysics, Vol. 4. 1966.
32. Williams, P. — Observatory, Vol. 86:67. 1966.

33. Dyson, F. — Science, Vol. 131:1667. 1960; Perspectives in Modern Physics, "Thoughts on the Search for Extra Terrestrial Technology." N.Y. Interscience Publishers. 1966.
34. Sagan, C. and R. Walker. — Ap. J., Vol. 144:1216. 1966.
35. Zel'dovich, Ya. B. — UFN. Vol. 89:647. 1966.
36. Rocchia, R., D. Rothenflug, D. Boclet, G. Dueros, and Y. Labeyrie. — 7th Int. Symp. Space Explor., Vienna, 11—17 May. 1966.*
37. Lebedev, D. S. and L. B. Levitin. Perenos informatsii elektromagnitnym polem (Information Transmission by Electromagnetic Fields). — In Sbornik: "Teoriya peredachi informatsii, Problemy peredachi informatsii," No. 16. 1964.
38. Vnezemnye Tsivilizatsii (Extraterrestrial Civilizations). Proceedings of a Conference, Byurakan, 20—23 May 1964. — Izd. AN Arm SSR. 1965.*
39. Kellerman, K. — Austr. J. Phys., Vol. 19:195. 1966.
40. Weaver, H., D. Williams, N. H. Dieter, and W. Lum. — Nature, Vol. 208:29. 1965.
41. Palmer, P. and B. Zuckerman. — HRAP, Vol. 124. 1966 (preprint).
42. Dieter, N. H., H. Weaver and D. Williams. — Sky and Telescope, Vol. 31:132. 1966.
43. Varshalovich, D. A. — ZhETF Letters, 4 (5):180. 1966.
44. Kaidanovskii, N. L. and N. A. Smirnova. — Radiotekhnika i Elektronika, Vol. 10:1574. 1965; Sky and Telescope, Vol. 34:143. 1967.
45. Symposium IAU, No. 29. Byurakan, May 1966.
46. Sholomitskii, G. B. — Astron. Zhurnal, Vol. 44:939. 1967.
47. Reddish, V. C. — Vistas in Astronomy, Vol. 7:173. 1966.

* [See footnote on p. 11.]

Chapter II

THE EFFECT OF THE SPACE MEDIUM ON THE PROPAGATION OF RADIO SIGNALS

The search for signals of extraterrestrial civilizations is closely associated with a painstaking analysis of the radio waves received from sources in outer space. The propagation of radio signals in the outer space is therefore one of the main topics in our analysis.

The outer space (including the interplanetary, the interstellar, and the intergalactic medium) is characterized by extremely low density of matter. The effect of the space medium on signal propagation is therefore also very low. However, because of the tremendous distances that the signals traverse before reaching the observer, the weak effects can build up to alarming magnitudes. The integrated cumulative effect may introduce significant distortions into the signal characteristics.

A detailed analysis of the propagation conditions clearly requires knowledge of the basic parameters of the space medium: density of matter, inhomogeneity of the medium, temperature, magnetic fields. These data (especially for the intergalactic medium) are fairly uncertain at this stage. Moreover, no detailed analysis can be carried out without giving consideration to the particular characteristics of certain limited regions of space through which the radio waves travel. For example, the conditions of propagation of radio waves in the Galactic plane are substantially different from the conditions of their propagation toward the Galactic pole. Individual objects (e.g., a dense cloud of ionized hydrogen) intercepting the line of sight may introduce significant distortions into the signal compared to the "average" propagation conditions.

We are unfortunately in a position to give only some general limiting estimates of the effect of the space medium on radio propagation.

Fairly numerous studies are available, dealing with such estimates. We will therefore consider the main conclusions pertaining to the "interference" from the space medium.

Absorption is one of the leading factors which affect the propagation of radio waves in a material medium. From the classical point of view, the absorption of radio waves can be described as oscillatory pumping of electrons by radio waves, which subsequently lose the extra energy through collisions with protons. The absorption coefficient per unit path length is expressed by the relation

$$\mu = \frac{1 - n^2}{cn} \nu_{\text{eff}}^{\text{coll}}, \qquad (2.1)$$

where n is the refractive index of the medium, ν_{eff}^{coll} is the effective number of electron-proton collisions. The expression for the refractive index can be written in the form

$$n(\omega) = \sqrt{1 - \frac{4\pi e^2 N_e}{m\omega^2}}. \qquad (2.2)$$

Here N_e is the electron concentration of the medium, ω is the frequency of the radio waves.

For the propagation of centimeter and decimeter waves in the rarefied interstellar and intergalactic medium, when $\frac{4\pi e^2 N_e}{m\omega^2} \ll 1$, the absorption coefficient is expressed in the form

$$\mu = \frac{0.58 N_e^2}{T^{3/2}\omega^2} \ln\left(\frac{4.3 \cdot 10^5 T}{\omega^{3/2}}\right), \qquad (2.3)$$

where T is the temperature of the medium.

The optical thickness for absorption (which measures the amount of absorption) $\tau = \mu l$ (l is the path length of the radio waves in the medium) is proportional to the measure of emission $N_e^2 \cdot l$, where N_e^2 is the square of the mean electron concentration in the medium along the entire path. The measure of emission for distances comparable with the size of the Galaxy lies between the limits $6 \cdot 10^{18} \leqslant N_e^2 l \leqslant 6 \cdot 10^{20}$.

The optical thicknesses τ for various measures of emission and frequencies ω are listed in Table 2.1.

TABLE 2.1. The optical thicknesses τ (in the Galaxy)

ω, Hz \ $N_e^{-2} \cdot l$	$6 \cdot 10^{20}$	$6 \cdot 10^{19}$	$6 \cdot 10^{18}$
10^{10}	$4 \cdot 10^{-5}$	$4 \cdot 10^{-6}$	$4 \cdot 10^{-7}$
10^{9}	$4 \cdot 10^{-3}$	$4 \cdot 10^{-4}$	$4 \cdot 10^{-5}$
10^{8}	$4 \cdot 10^{-1}$	$4 \cdot 10^{-2}$	$4 \cdot 10^{-3}$
10^{7}	40	4	0,4

The table shows that radio waves with frequencies $\omega \geqslant 10^9$ Hz propagate virtually without absorption in any direction in the Galaxy. (The magnitude of absorption is determined by the factor $e^{-\tau}$.) Transmission at lower frequencies is obstructed by strong absorption, especially in the direction of the Galactic plane, where $N_e^2 \cdot l$ reaches its maximum values.

The refractive index (2.2) also determines the dispersion effects which distort the transmission. We should distinguish between two effects: the phase shift of the spectral components of the signal due to dispersion in the medium and the "lag" of the quasimonochromatic group components which transmit the signal energy.

To illustrate the difference between these two effects, let us consider the transmission through space of a train of pulses, pulses of length P following one another at intervals of the same length (Figure 28).

A pulse of length P (and any train of such pulses) can be expanded into a spectrum (a Fourier integral). The train of pulses shown in Figure 28 has a spectrum which covers a frequency interval of width $\Delta\omega \sim \frac{2\pi}{P}$.

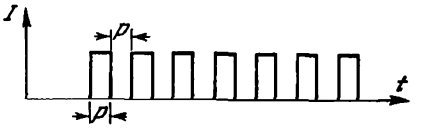


FIGURE 28. Undistorted pulses.

In dispersive media ($n \neq 1$), the phases of the individual spectral components of the signal propagate with different velocities $v_{ph} = \frac{c}{n(\omega)}$. Therefore, the phases of different components acquire a relative shift and the resultant combination at the receiver gives a distorted pulse shape /1/. Depending on the characteristics of the propagating medium, the pulse either "contracts" or "spreads."

We have considered the propagation of high-frequency pulses. Each pulse was "filled" with monochromatic radiation of a high frequency ω. This treatment is valid only for a steady-state transmission of a high-frequency signal (when $\omega P \gg 1$). This means that the pulse length accommodates a considerable number of periods of oscillations of frequency ω. In the opposite case ($\omega P < 1$) the pulse involves a macroscopic variation of the intensity of a low-frequency field in the medium.

Suppose that the pulse train (see Figure 28) is generated in the following way: a certain radiation source with a sufficiently wide continuous spectrum is periodically obscured by a screen. The pulse train emitted into space will then have a wide-band "filling," possibly not unlike noise (thermal "noise").

Let the frequency band of this radiation be $\Delta\omega$, and the spectral density $E(\omega)$. From the energy point of view, each pulse is a collection of quasi-monochromatic groups $E(\omega)\delta\omega$, where $\delta\omega$ is a very narrow "quasimonochromatic" band in the spectrum. The integrated effect of all these group intensities gives the height and the length of the pulse. Quasimonochromatic wave groups propagate with a group velocity $v_{gr} = cn(\omega)$. For ionized space media, this velocity decreases with decreasing frequency. As a result, over sufficiently long distances, the high-frequency wave groups precede the low-frequency groups, and a characteristic "time sweep" of the spectrum is obtained. The importance of this effect in connection with solar radio bursts with frequency drift was discussed in /6/. A similar effect relating to the propagation of radio waves in interstellar media was discussed in /2/. The problem was also considered in /3, 4/.*

* This effect was first discovered in observations of pulsars — pulsating radio sources.

The delay of a wave group of frequency ω_1 relative to a wave group of frequency ω_2 can be found from the relation

$$\Delta t(\omega_1 - \omega_2) = \frac{l}{cn(\omega_1)} - \frac{l}{cn(\omega_2)}. \tag{2.4}$$

If the frequencies ω_1, ω_2 are far from the critical frequency ω_{cr} at which $n(\omega_{cr}) = 0$ (this condition can be written in the form $n \ll 1$), the delay is expressed by the formula

$$\Delta t(\omega_1 - \omega_2) \sim \frac{2\pi e^2}{mc} \frac{\omega_1^2 - \omega_2^2}{\omega_1^2 \omega_2^2} \bar{N}_e \cdot l. \tag{2.5}$$

This expression is conveniently rewritten taking ω_1 and ω_2 in the form

$$\omega_1 = \omega_0 + \frac{\Delta\omega}{2},$$

$$\omega_2 = \omega_0 - \frac{\Delta\omega}{2},$$

where ω_0 is the mean frequency of the signal. Then

$$\Delta t(\Delta\omega) = \frac{2\pi e^2}{mc} \frac{\Delta\omega}{\omega_0^3} N_e \cdot l. \tag{2.6}$$

The delay Δt for space media for various $\Delta\omega$ and ω_0 is listed in Table 2.2. We see from the table that the delay may reach considerable values. What does this lead to ?

To answer this question, we have to consider the conditions of reception of the signal shown in Figure 28. Suppose the receiver band $\Delta\omega_{rec} \geqslant \Delta\omega$, i. e., the receiver is capable of receiving the full intensity of the entire spectrum of the signal $\Delta\omega$. For simplicity, we take $E(\omega) = \text{const}$ in the entire frequency band $\Delta\omega$. Clearly, if $\Delta t(\Delta\omega) \ll P$, no significant distortions will be introduced in the received signal. If, however, $\Delta t \sim P$ (Figure 29,a) the signal is markedly distorted. The high-frequency spectral components of the signal are the first to be received. The low-frequency components are delayed and the signal "spreads."

If $\Delta t \gg P$ (Figure 29, b), the pulses are completely blurred into a continuous emission from the source (its intensity is much less than the peak pulse power).

The true signal shape in principle can be restored by an appropriate correction in the receiver or in the processing stage. The unfortunate fact, however, is that we do not have the actual numerical values of the parameters of the media propagating the pulses from outer space.

The periodicity can be "caught" (for $\Delta t \gg P$) by narrowing the receiver band $\Delta\omega_{rec}$ to such an extent that $\Delta t(\Delta\omega_{rec}) < P$. This procedure, however, will lead to substantial losses of the receiver sensitivity (which is proportional to $\left(\frac{\Delta\omega_{rec}}{\Delta\omega}\right)^2$, a factor not to be trifled with in the reception of signals from outer space. Moreover, the narrower receiver band will have an adverse effect on the rate of information transmission (see Chapter III).

TABLE 2.2. Lag time for wave groups, in sec (Galaxy, Metagalaxy)

$\bar{N}_e \cdot l$ / ω_0, Hz	$\Delta\omega = 0.5\omega_0$	$\Delta\omega = 0.1\omega_0$	$\Delta\omega = 10^{-3}\omega_0$	$\Delta\omega = 10^{-5}\omega_0$
	$5 \cdot 10^{20}$ (the limit for interstellar distances)			
10^{10}	0,25	$5 \cdot 10^{-2}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-6}$
10^{9}	25	5	$5 \cdot 10^{-2}$	$5 \cdot 10^{-4}$
10^{8}	$2{,}5 \cdot 10^{3}$	500	5	$5 \cdot 10^{-2}$
10^{7}	$2{,}5 \cdot 10^{5}$	$5 \cdot 10^{4}$	$5 \cdot 10^{2}$	5
	$5 \cdot 10^{21}$			
10^{10}	2.5	0.5	$5 \cdot 10^{-3}$	$5 \cdot 10^{-5}$
10^{9}	250	50	0.5	$5 \cdot 10^{-3}$
10^{8}	$2.5 \cdot 10^{4}$	$5 \cdot 10^{3}$	50	0.5
10^{7}	$2.5 \cdot 10^{6}$	$5 \cdot 10^{5}$	$5 \cdot 10^{3}$	50
	$2 \cdot 10^{23}$ (the limit for intergalactic distances)			
10^{10}	10^{2}	0.2	$2 \cdot 10^{-3}$	$2 \cdot 10^{-5}$
10^{9}	10^{4}	20	0,2	$2 \cdot 10^{-3}$
10^{8}	10^{6}	$2 \cdot 10^{4}$	20	0.2
10^{7}	10^{8}	$2 \cdot 10^{6}$	$2 \cdot 10^{4}$	20

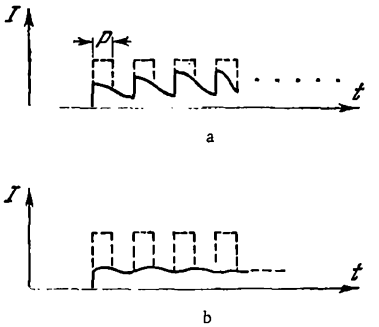


FIGURE 29. Distorted pulses:

a) the case $\Delta t \sim P$; b) the case $\Delta t \gg P$.

TABLE 2.3. Minimum pulse length (seconds)

$\bar{N}_e \cdot l$ / ω, Hz	$5 \cdot 10^{20}$	$5 \cdot 10^{21}$	$5 \cdot 10^{22}$	$2 \cdot 10^{23}$
10^{10}	$2.5 \cdot 10^{-5}$	$8 \cdot 10^{-5}$	$2.5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
10^{9}	$8 \cdot 10^{-4}$	$2.5 \cdot 10^{-3}$	$8 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$
10^{8}	$2.5 \cdot 10^{-2}$	$8 \cdot 10^{-2}$	0.25	0,5
10^{7}	$8 \cdot 10^{-2}$	2,5	8	15

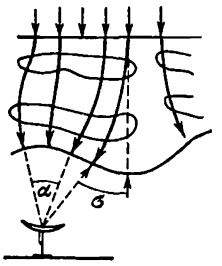


FIGURE 30. Wave front distortion.

The group lag effect imposes certain restrictions on the permissible pulse length P. To avoid the highly undesirable significant distortions which we described above, we have to ensure the inequality $\Delta t(\Delta\omega) < P$. The minimum values of P prescribed by this requirement are listed in Table 2.3.

Radio waves propagating over large distances in the intergalactic medium may also show a "red shift." The red shift has an "unfavorable" effect, lowering the frequency of the propagating radiation. The distortion effects are therefore enhanced for propagation over very large distances.

The distortions introduced by the space medium into other types of radio signals (frequency or phase modulated signals) should be considered separately.

Analysis of the data in Tables 2.2 – 2.3 stresses the advisability of using the shortest wavelengths in the radio spectrum for long-range interstellar communication.

The effect of radio wave propagation conditions in the space medium and in the Earth's atmosphere on the apparent angular size of the radio source has been discussed in /5/. If the propagating medium is inhomogeneous, the wave front is distorted on passing through this medium (Figure 30). The amount of wave front distortion is determined by the deviation of the wave phase from the unperturbed value. Proceeding from some model considerations (e.g., the size of inhomogeneities, the mechanism of wave scattering by the inhomogeneities, etc.), we can arrive at an average statistical estimate of the integrated distortion acquired by a wave on passing through an inhomogeneous layer of a given thickness. The mean square phase deviation from the unperturbed value, $\bar{\varphi}^2$, was calculated in /5/ using the expression

$$\bar{\varphi}^2 = \frac{4\pi^2 l \cdot d}{\lambda^2} \overline{\Delta n^2}, \qquad (2.7)$$

where l is the path length of the wave in the inhomogeneous medium, d is the mean inhomogeneity size (it is assumed that $d \gg \lambda$), λ is the wavelength, $\overline{\Delta n^2}$ is the mean square fluctuation in the refractive index of the medium. Using the geometrical optics approximation, we can obtain an expression for the deviation angle σ of the beam from the original source – observer direction. We have

$$\bar{\sigma}^2 = 4\pi^{1/2} \frac{l}{d} \overline{\Delta n^2}. \qquad (2.8)$$

The spreading of the angular diameter of the source to σ will be observed in the far zone of the scattering region (i.e., at distances of the order $R \approx l \geqslant \frac{d^2}{\lambda}$). Tables 2.4 and 2.5 show that for $\bar{\varphi}^2 \gg 1$, this condition is not satisfied on the Earth. The Earth-bound observer is located in the near scattering zone of the space inhomogeneities. In this case, the distortions introduced by the medium in the size of the source are determined by the

relations between $\sqrt{\overline{\sigma^2}}$, the angular size of the inhomogeneities $\sqrt{\overline{\alpha^2}}\left(\alpha \approx \frac{d}{l}\right)$, and the antenna size D.

The values of $\sqrt{\overline{\sigma^2}}$, $\sqrt{\overline{\alpha^2}}$ for various media are listed in Table 2.5.

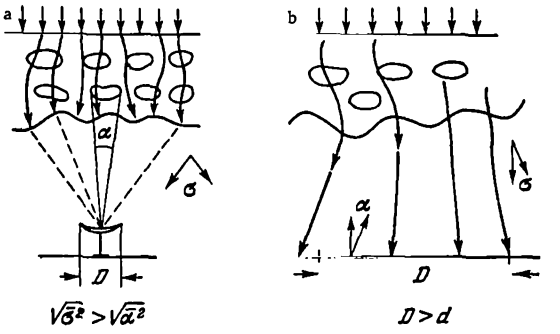


FIGURE 31. Deviation of light rays propagating through an inhomogeneous medium.

If $\sqrt{\overline{\sigma^2}} > \sqrt{\overline{\alpha^2}}$, an antenna for any diameter will receive rays which have covered a distance greater than the correlation radius (as determined by the mean inhomogeneity size), and the source will expand to the full angular size σ (Figure 31, a).

TABLE 2.4. *The parameters of space media used for the calculations in Table 2.5*

Medium		l, cm	d, cm	$\sqrt{\overline{\Delta n^2}}$	$\overline{N}_e$
Troposphere		$1.5 \cdot 10^6$	$6 \cdot 10^3$	$0.5 \cdot 10^{-6}$	–
Ionosphere		$4 \cdot 10^7$	$2 \cdot 10^4$	$4.5 \cdot 10^{-12}\lambda^2$	
Interplanetary	Ecliptic plane	10^{14}	10^9	$4.5 \cdot 10^{-12}\lambda^2$	10^2
	Toward the pole	$0.5 \cdot 10^{13}$	10^9	$0.5 \cdot 10^{-12}\lambda^2$	20
Interstellar	Galactic plane	6.10^{22}	$3 \cdot 10^{18}$	$4.5 \cdot 10^{-14}\lambda^2$	$3 \cdot 10^{-2}$
	To the Galactic pole	$6 \cdot 10^{20}$	$3 \cdot 10^{18}$	$4.5 \cdot 10^{-14}\lambda^2$	$3 \cdot 10^{-2}$
Intergalactic		10^{28}	10^{22}	$4.5 \cdot 10^{-19}\lambda^2$	10^{-5}

If $\sqrt{\overline{\sigma^2}} \ll \sqrt{\overline{\alpha^2}}$, the effect will vary depending on the relation between D and d.

In a filled-aperture antenna of size $D > d$ (Figure 31, b) the phase fluctuations produced by inhomogeneities cause a loss in the effective area and broaden the beam angle.

For antennas of size $D \ll d$ (Figure 30), refraction effects are observed, which shift the apparent position of the source through the refraction angle. As a result, the source coordinates are measured with a certain error.

TABLE 2.5. Distortion of point source image

Space medium		Φ^2 rad^2	$\sqrt{\bar{\sigma}^2}$, rad	$\sqrt{\bar{\alpha}^2}$	Range of wavelengths where $\varphi^2 > 1$, $\sqrt{\bar{\sigma}^2} \ll \sqrt{\bar{\alpha}^2}$	
					$\varphi^2 > 1$	$\sqrt{\bar{\sigma}^2} \ll \sqrt{\bar{\alpha}^2}$
Interplanetary medium	Ecliptic plane	$1.4 \cdot 10^2 \lambda^2$	$4 \cdot 10^{-9} \lambda^2$	10^{-5}	entire spectrum	$\lambda < 16$ cm
	Polar	$0.28 \lambda^2$	$1.7 \cdot 10^{-10} \lambda^2$	$2 \cdot 10^{-4}$	$\lambda > 2$ cm	$\lambda < 3.5 \cdot 10^2$ cm
Interstellar medium	Galactic plane	$2.5 \cdot 10^{16} \lambda^2$	$1.7 \cdot 10^{-11} \lambda^2$	$5 \cdot 10^{-5}$	entire spectrum	$\lambda < 5.5 \cdot 10^2$ cm
	Polar	$2.5 \cdot 10^{14} \lambda^2$	$1.7 \cdot 10^{-12} \lambda^2$	$5 \cdot 10^{-3}$	entire spectrum	entire spectrum
Intergalactic medium		$1.4 \cdot 10^{15} \lambda^2$	$1.2 \cdot 10^{-15} \lambda^2$	10^{-6}	entire spectrum	entire spectrum

In these calculations, the possible motion of the inhomogeneities should be taken into consideration. If the inhomogeneity clouds move with a certain velocity v, the source will "shimmer" with a period

$$t = \frac{d}{v}. \tag{2.9}$$

Scattering effects thus limit the resolving power of antennas. There are two alternatives: either the source expands ($\sqrt{\bar{\sigma}^2} > \sqrt{\bar{\alpha}^2}$) or the scattering has an adverse effect on antenna directivity ($\sqrt{\bar{\sigma}^2} \ll \sqrt{\bar{\alpha}^2}$, $D > d$).

Table 2.5 lists the wavelength region for the various space media where the conditions $\varphi^2 \gg 1$, $\sqrt{\bar{\sigma}^2} \ll \sqrt{\bar{\alpha}^2}$ are satisfied. The interplanetary medium evidently introduces considerable distortion in the angular size of the source.

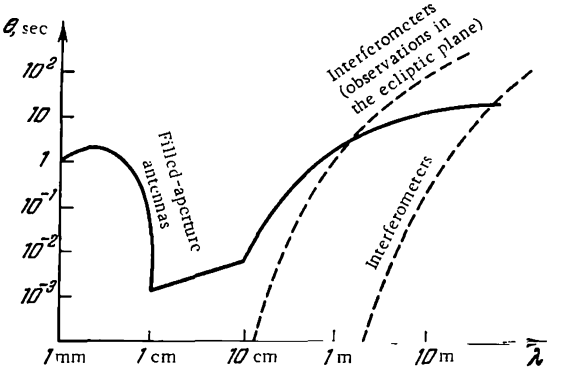


FIGURE 32. Limiting antenna resolution.

Figure 32 plots curves of the limiting antenna resolution derived with allowance for the effect of scattering in space media. Filled-aperture antennas are very limited in terms of resolution. The maximum resolving power is attainable only using radio interferometers (at wavelengths

shorter than 10 cm). This again stresses the advisability of using the shortest wavelengths of the radio spectrum in observations.

In conclusion note that the "reversal" of the problem of distortions introduced by the space medium may prove quite fruitful for astrophysical purposes. If the true dimensions of the source or the parameters of the variable radio signal from the source can be estimated from independent considerations, the analysis of distortions introduced by the space medium may provide highly valuable information on the properties of the medium itself (material density, size of inhomogeneities, etc.).

Bibliography

1. Ginzburg, V.L. Rasprostranenie elektromagnitnykh voln v plazme (Propagation of Electromagnetic Waves in Plasma). – Fizmatgiz. 1960.
2. Gudzenko, L.I. and B.N. Panovkin. – In: "Vnezemnye tsivilizatsii," Proceedings of a Conference. Byurakan, 20–23 May 1964, p. 68. Izd. AN Arm. SSR, 1965.*
3. Haddock, F.I. and D.W. Sciana. – Phys. Sci. Let., 14(25):1007. 1965.
4. Panovkin, B.N. – Fifth Soviet Conf. on Radio Astronomy, Khar'kov. 1965.
5. Kaidanovskii, N.L. and N.A. Smirnova. – Radiotekhnika i Elektronika, Vol. 10:1574. 1965.
6. Wild, J.P., K.V. Sheridan, and A.A. Neylan. – Austr. J. Phys., Vol. 12:369. 1959.

* [See footnote on p. 11.]

Chapter III

THE POSSIBILITY OF RADIO COMMUNICATION WITH EXTRATERRESTRIAL CIVILIZATIONS

The topic of communication with extraterrestrial civilizations (EC) has repeatedly cropped up in the scientific literature /2, 3/ after the pioneering work of Cocconi and Morrison /1/, who were the first to establish the feasibility of communication with EC in the electromagnetic spectrum. There is no doubt that the organization of communication with EC is an unprecedented technical problem, whose specific requirements cannot be fully appraised at this stage. On the other hand, it seems that any communication system, including the system of communication with EC, would satisfy certain general requirements which follow from the general laws of information transmission. The study of these laws is the subject of the information theory or the general theory of communication. We will therefore start our review with a discussion of the principal elements of the general theory of communication, which will prove useful in the following.

§1. ELEMENTS OF THE GENERAL THEORY OF COMMUNICATION

Structure and fundamental characteristics of a communication system

The aim of any communication system is the transmission of certain messages. The messages may constitute text written using the letters of a certain alphabet (as in telegraph messages) or sounded verbally (telephone, radio). The message may also constitute an image of a certain object (phototelegraph, television) or an algorithm to be transmitted to an automatic control system. Any of these messages can be represented as a succession of digits or as some continuous time function $x(t)$.

Messages are transferred by a communication system using certain agreed signals. In the present chapter we will only consider systems employing electrical signals.* An electrical signal is a time-variable

* In a more general treatment of communication, when we are dealing with such systems as biological population, biological evolution, etc., the basic concepts of message, signal, and information require a new, more precise definition. However, after an appropriate generalization of these concepts, the basic propositions of the theory of electrical communication prove to be valid for a larger class of communication systems.

electrical magnitude (voltage, current, field strength) and, like the message itself, it can be expressed as a certain function of time. The signal reflects the message in the form of an electrical disturbance. The transmitted message must be reconstituted from the received signal.

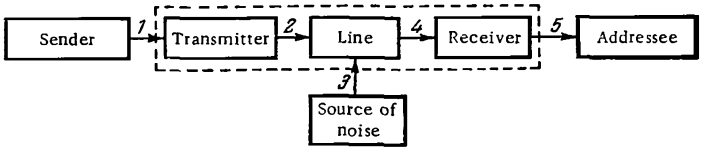


FIGURE 33. A generalized communication system:

1 — message, 2 — signal, 3 — noise, 4 — signal + noise 5 — received message. The part of the block diagram enclosed in the dashed rectangle is the communication channel.

A block diagram of a generalized communication system is shown in Figure 33. The message from the information source, or the sender, is delivered to the transmitter which transforms it into a signal sent through the communication line. The communication line is an electromagnetic wave channel, or, in other words, the medium propagating the signal from the transmitting end of the system to the receiving end. The line may comprise two conducting wires, a coaxial cable, a waveguide, or the unrestricted part of space in which radio waves propagate. Thus, for short-wave radio communication, this is the spherical layer between the Earth's surface and the ionosphere. In directional radio transmission, the communication line is the part of space inside the solid angle subtended by the receiving antenna.

A signal propagating along the communication line may experience distortion and may be intermixed with noise. Distortion is generally described as those changes in the signal which are caused by known characteristics of the system. In principle, these distortions can be corrected, and we will not have to analyze their effects.* Noise, on the other hand, is random and cannot be fully corrected. Random noise is of the greatest importance for the actual performance of a communication line.

At the receiving end of the line, the electrical message is picked up by a receiver, which constitutes the original message by an appropriate transformation of the received signal. Mathematically, the action of the receiver is the inversion of the transmitter action. The part of the system including the transmitter, the line, and the receiver is generally designated as the communication channel.

* Distortions experienced by a signal propagating in the interstellar medium were considered in Chapter II. Note that when propagating in a medium with randomly changing properties, the signal experiences random distortions which cannot be corrected. An example of such random distortions is the scintillation of stars and radio sources. The same effect causes a definite broadening of the angular dimensions of the sources (see Chapter II).

In an ideal communication system, free from noise, the received message is identical to the transmitted message. In real noisy systems, however, this is never so. The degree of identity of the received and transmitted signals characterizes the reliability of the communication system. The reliability depends on the ratio of the signal power to the noise power in the communication signal. As a rule, the reliability falls off with distance. The maximum distance over which a certain reliability is still attainable is known as the communication range. This parameter is naturally of the greatest importance in systems of communication with EC.

Another important characteristic is the transmission rate of the communication channel, i.e., the quantity of information that can be transmitted by the given communication system in unit time. The transmission rate characterizes the information content of the transmitted message. However, the system does not "distinguish" between important and trivial messages. Thus, to send a telegram consisting of 100 symbols, the system should always meet certain fixed requirements (transmission time, frequency band, signal power, etc.), regardless of the importance and the content of the message. The concept of information in the general theory of communication is therefore devoid of any qualitative meaning, and should be treated as a pure quantitative concept.

Quantitative definition of information

How are we to define information? Consider the transmission of a sequence of four-digit decimal numbers. These are either numerical values of some physical magnitude or four-letter words written using a ten-letter alphabet. Suppose we are transmitting a certain word M. What is the information content of our message? The total number of four-digit numbers or possible messages is $N = 10^4$. By transmitting our message, i.e., a particular number M, we have made a definite choice out of the available total of $N = 10^4$. The number N of the available choices characterizes the uncertainty of the outcome prior to the transmission. This number N is also used to characterize the information content of the particular message. The higher the initial uncertainty which prevailed before the transmission, the higher is the quantity of information contained in the message, and conversely: the lower the initial uncertainty, the lower is the quantity of information in the transmitted message. If the quantity of information is designated Q, we may write

$$Q = Q(N), \tag{3.1}$$

where Q is a single-valued monotonically increasing function of N. From this definition we see that if $N_1 = N_2$, then

$$Q_1 \equiv Q(N_1) = Q(N_2) = Q_2. \tag{3.2}$$

The four-digit decimal number M can be expressed in a binary, ternary, or any other number system with some base a. Then $N_1 = a_1^{m_1}$, $N_2 = a_2^{m_2}$, where m is the exponent of the number M (for a whole m, this is simply the number of digits needed to express the number in the given system). Condition (3.1) thus takes the form

$$Q_1 = Q_2, \quad \text{if} \quad a_1^{m_1} = a_2^{m_2}. \tag{3.3}$$

In this form, it simply means that the quantity of information contained in the number M is independent of the particular system used to express this number.

The next condition to be met by our definition of information is that if we take different numbers expressed in the same number system, the information content of each number will be proportional to the number of digits (i. e., a six-digit decimal number 145876 contains double the information of the three-digit number 963 and three times as much information as the two-digit number 25). Thus,

$$Q = \gamma m, \tag{3.4}$$

where γ is a proportionality coefficient.

In application to the problem of information transmission through a channel, this means that the quantity of transmitted information increases linearly with transmission time (indeed, to transmit six digits, we need double the time to transmit three digits). Thus, a two-minute transmission is in general (other conditions being equal) more informative than a one-minute transmission.

It can be shown that conditions (3.1) – (3.4)* define a unique function

$$Q = \log_b N = m \log_b a. \tag{3.5}$$

This definition was first advanced by Hartley /4/ in 1928 and it has been used since with excellent results in the theory of communication.

The base b of the logarithm in (3.5) is arbitrary. The choice of this base corresponds to the unit of information measurement. Taking $b = 2$, we obtain the quantity of information Q in binary units, or bits. This unit of information, corresponding to the lowest possible base of a number system, may be adopted as the basic unit of information measurement. It is widely used in applications.

Let us now determine the information content of our four-digit decimal number M. Taking $m = 4$ and $a = 10$, we find $Q = 4 \log_2 10 \approx 13.3$ bits. Similarly, the quantity of information in a five-letter word from a 30-letter alphabet is $5 \log_2 30 = 24.6$ bits, and a text of 100 words with an average word length of 5 letters contains about 2460 bits of information. For $a = b = 2$, $Q = m$ bits, i. e., the quantity of information, expressed in bits, contained in a number M is equal to the number of binary digits required to express this number in a system with a base 2 (for whole m, naturally).**

* Since conditions (3.1) – (3.3) are not independent, any two of the conditions are sufficient for a single-valued definition of Q, e. g., (3.1) and (3.4), (3.2) and (3.4), or (3.3) and (3.4).

** If m is not a whole number, M is expressed using m_1 binary digits, where m_1 is the nearest whole number to m, $m_1 > m = Q$.

The above definition applied to discrete messages. However, a continuous function of time can be represented with any desired accuracy by a set of discrete quantities, and this definition is therefore quite general for the purposes of communication theory.

Let us now consider the various techniques whereby a message is transformed into a signal.

Transformation of a message into a signal. Forms of modulation

A signal is transmitted as a direct current, electromagnetic oscillations of high frequency, or a periodic train of pulses. When a signal is transmitted down a communication line, one of the line parameters varies in accordance with the transmission function $x(t)$.

Direct current is characterized by two parameters: the magnitude and the direction of the current. By changing one of these magnitudes in accordance with $x(t)$, we obtain an electric signal which may propagate along the communication line (e.g., as in the transmission of Morse-coded telegrams). However, since direct current will propagate only through wires, this method of transmission is of no consequence for our problem.

The signals in radio communication are high-frequency electromagnetic oscillations which may propagate freely through the vacuum. Sinusoidal oscillations are characterized by three parameters: the amplitude, the frequency, and the initial phase. By altering one of these parameters in accordance with the message function, often called the modulating function, we obtain a modulated electromagnetic signal of high carrier frequency. We thus distinguish between three different forms of modulation, corresponding to the three parameters of the carrier: amplitude modulation AM, frequency modulation FM, and phase modulation PM (Figure 34a). The modulated signals are demodulated in the receiver to reconstitute the modulating function $x(t)$, which is the message.

If the signal is transmitted by a periodic train of pulses, we obtain four types of pulse modulation corresponding to variation of the pulse height h, pulse duration τ, and pulse recurrence frequency $\nu_0 = \frac{1}{T}$ (T is the time between two successive pulses): these are the pulse-amplitude modulation, PAM, the pulse-duration modulation PDM, the pulse frequency modulation PFM, and the pulse position modulation PPM (Figure 34b). In a number of cases repeated modulation is used: the pulse train is modulated by the message function, and the modulated pulses are then used to modulate a high-frequency carrier (Figure 34c). This provides a new modulation technique, high-frequency pulse modulation HFPM, in which the height, length, frequency, and phase of pulses remain constant, and only the duty cycle is altered (Figure 34d).

An important variety of pulse modulation is the transmission of coded messages. We will consider this type of modulation after becoming better acquainted with some properties of signals.

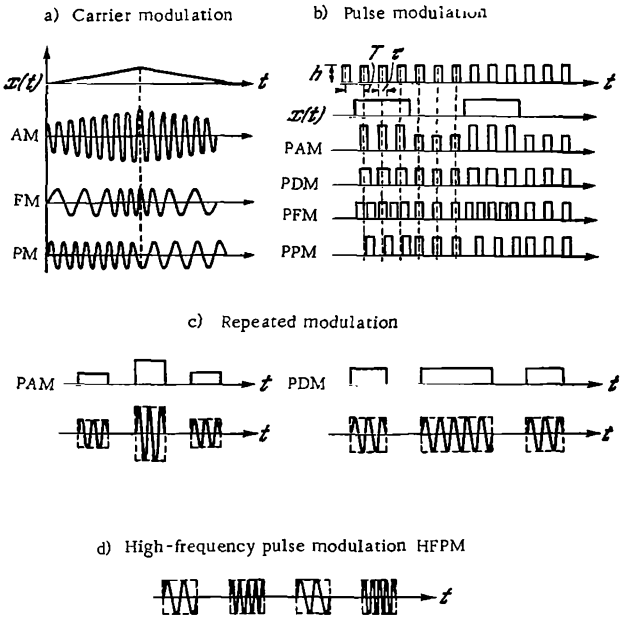


FIGURE 34. Formation of electrical signals by modulation:

$x(t)$ is the message function or the modulating function, AM amplitude modulation, FM frequency modulation, PM phase modulation; PAM pulse-amplitude modulation, PDM pulse-duration modulation, PFM pulse-frequency modulation, PPM pulse position modulation.

Physical characteristics of signals

A signal can be characterized by the following three parameters: signal duration, the dynamic range, and band width.

Signal duration is the simplest characteristic. Its practical importance is self-evident: the longer the signal, the longer it takes to transmit it and the longer the lines remains engaged.

The dynamic range is defined as the ratio of the maximum instantaneous signal power (the so-called peak power) to the minimum signal power. The dynamic range is measured on a logarithmic scale and is expressed in decibel. One decibel (1 dB) is equal to 0.1 on the logarithmic scale; therefore if n is the ratio of the measured quantities on the linear scale, the same ratio in dB is equal to $10 \log n$. Signals where the peak power is double the minimum power have a dynamic range of 3 dB; a dynamic range of 10 dB corresponds to a maximum-to-minimum ratio of 10, 20 dB to a ratio of 100, 30 dB to a ratio of 1000, etc.

The choice of the minimum signal power is determined by the noise level. To ensure a reliable reception, the minimum signal power should exceed by a certain factor the mean noise power P_n ($P_{min}=\alpha P_n$). High-quality transmission of speech by amplitude modulation requires P_{min} exceeding the mean noise power by 60 — 70 dB. The quantity $P_{min}=\alpha P_n$

is known as the threshold signal power. The dynamic range, related to the threshold power, is often replaced by the ratio of the mean signal power to the mean noise power P_s/P_n. This ratio is briefly called signal-to-noise ratio or signal/noise ratio. Both the dynamic range and the signal-to-noise ratio characterize the signal power relative to the noise power, and not the absolute power. What are the factors determining the threshold power?

Suppose we wish to transmit a certain message, which expresses the value of the function $x(t)$ at the time t_0. We may use one of the pulse modulation systems, e.g., the pulse-amplitude modulation, and send a pulse of height $x(t_0)=x_0$ along the communication line. In the case of an ideal noise-free channel, this pulse is received without distortion at the receiving end of the communication line, and the original message $x(t_0)$ will be recovered from the pulse amplitude x_0. In a real channel, the signal is mixed with noise, and the received pulse amplitude is therefore $x_0+\xi$, where ξ is the noise amplitude (positive or negative). Suppose we are interested in recovering the message with an accuracy of 0.001, i.e., the relative error is $\frac{\Delta x_0}{x_0}=0.001$. To this end, we should have

$$|\xi|<\frac{1}{2}\Delta x_0. \tag{3.6}$$

If $|\xi|=\text{const}$, i.e., the noise is constant, this condition is satisfied when $x_0>2000\,|\xi|$ or $P_s=x_0^2>4\cdot 10^6\xi^2=4\cdot 10^6 P_n$. In other words, the signal must be a factor of four million more powerful than the noise level (66 dB). This is the threshold signal power for the PAM transmission of the instantaneous value of $x(t)$ with an error not exceeding 0.001.

This case of constant-noise communication is trivial: constant noises are easily corrected. The main difficulty is that the real noise is a random function which cannot be corrected. Random noise, in general, may take on arbitrarily large values, although the probability of this event is low. To determine the threshold power in the presence of random noise, we have to find the probability that the noise does not exceed $\frac{1}{2}\Delta x_0$, i.e., the probability that condition (3.6) is satisfied. If we are dealing with Gaussian noise, i.e., noise with normally distributed amplitudes, the sought probability is

$$p_0\left(|\xi|<\frac{1}{2}\Delta x_0\right)=\Phi\left(\frac{\Delta x_0}{2\sqrt{2}\,\sigma}\right)=\Phi(z), \tag{3.7}$$

and the probability of error is

$$p=1-p_0=1-\Phi(z), \tag{3.8}$$

where σ is the parameter of the Gaussian distribution, $\sigma=\sqrt{\xi^2}=\sqrt{P_n}$ and Φ is the Laplace function, or the probability integral. This integral has been tabulated in detail, and the sought probability can be extracted from the corresponding table. For $\Delta x_0=10\sigma$, the probability of error is of the order of 10^{-6}, and then it falls off rapidly as $\frac{\Delta x_0}{\sigma}$ increases. For most practical problems, the reliability corresponding to an error probability of 10^{-6} is quite sufficient. We can thus ensure reliable transmission (in the above

sense) with signal reproducibility of $\frac{\Delta x_0}{x_0} = 0.001$ if $x_0 = 1000\,\Delta x_0 = 10^4\sigma$ and $\frac{P_s}{P_n} = 10^8$.

We have considered the determination of threshold power in the simplest case of message transmission by a single pulse. The results, however, remain valid for any complex electrical signal $x(t)$. In the general case, x_0 is to be interpreted as the minimum signal amplitude ($x_0^2 = P_{min}$).

Note that for a given dynamic range and given minimum signal power, the minimum power is also well determined. As the mean power is reduced, the communication becomes unreliable. Thus, besides the minimum threshold power $P_{min} = \alpha P_n$, we can also speak of the threshold mean power of the signal. Later, when dealing with the transmission of continuous functions by pulsed signals, we will show how to determine the threshold mean power for certain types of signals (PCM with an arbitrary code base). Now we will consider the spectral characteristics of a signal.

Any periodic function $x(t)$ of period T can be written as a sum of harmonic vibrations of multiple frequencies (a Fourier expansion):

$$x(t) = \sum_{k=1}^{\infty} c_k \cos(\omega_k t + \varphi_k). \tag{3.9}$$

Each component (harmonic) of this expansion is a sinusoidal vibration of frequency ω_k, amplitude c_k, and phase φ_k. The frequencies of the individual harmonics are integral multiples, and are related to the period of the function by the equality $\omega_k = k\frac{2\pi}{T}$ $(k = 1, 2, 3, \ldots)$. The lowest frequency is $\omega_1 = \frac{2\pi}{T}$, and this is also the difference between the frequencies of any two successive harmonics. The values of c_k and φ_k depend on the form of the function $x(t)$. The set of the coefficients c_k form the amplitude spectrum, and φ_k the phase spectrum. Such a spectrum, consisting of individual discrete values, is known as a line spectrum. As the period increases, the spacing between the lines decreases, and in the limit for $T \to \infty$ (i.e., a nonperiodic function), we obtain a continuous spectrum (Figure 35). Mathematically, a continuous spectrum is expressed by a Fourier integral.

Knowledge of the amplitude spectrum and the phase spectrum completely defines the function $x(t)$. Therefore, any process may be described either by defining the appropriate time function or by specifying the spectrum, which is a function of frequency. Both the time and the frequency representations are equivalent.

All the signals encountered in practice are bounded-spectrum functions. This means that they do not contain frequencies below some minimum frequency ν_1 and above some maximum frequency ν_2. They occupy a finite frequency band from ν_1 to ν_2. The band of frequencies filled by the spectrum of the signal defines the signal band width $\Delta\nu = \nu_2 - \nu_1$. This is a highly important characteristic of the signal. In transmission along a communication channel, the signal frequency band may shift toward higher or lower frequencies in the spectrum.

However, the band width $\Delta\nu$ remains unchanged by this shift.* The frequency shift is very useful in radio engineering, e.g., in superheterodyne receivers. The application of this effect in communication systems makes possible simultaneous transmission of numerous messages along a single communication line, by using different frequencies.

The greater the band width Δf of the communication line, the higher is the number of signals with a given band width $\Delta\nu$ that can be transmitted simultaneously. Each signal is associated with a certain message, characterized by a definite quantity of information. We thus conclude that the rate of information transmission through a certain communication channel is proportional to the channel band width.

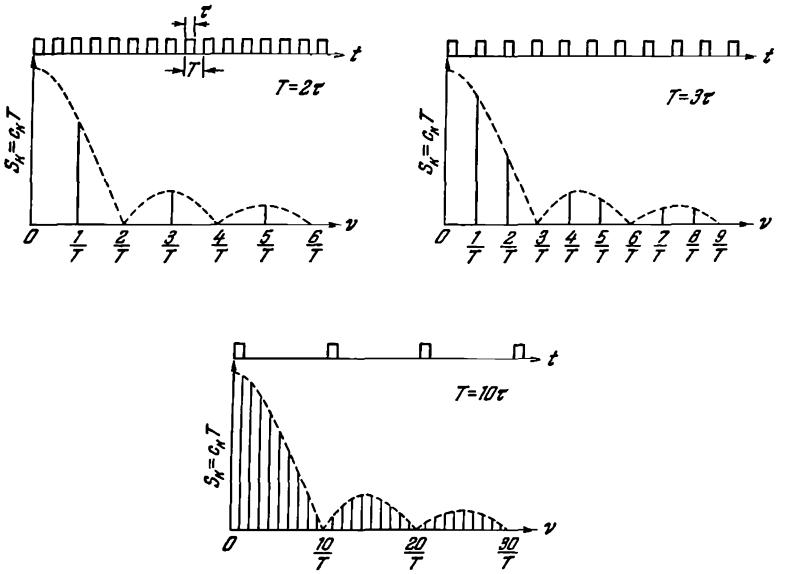


FIGURE 35. The spectrum of a periodic pulse train.

The vertical axis gives $S_k = c_k T$ (the product of the amplitude of the corresponding harmonic and the period). The dashed line gives the spectral density of the amplitude of a unit pulse. As the period is increased, the spacing between the spectral lines diminishes and in the limit $T \to \infty$ a continuous spectrum is obtained, which coincides with the spectrum of a unit pulse.

* In certain stages of the transmission process, the signal band width $\Delta\nu$ may indeed change. Thus, in FM, the band width of the signal in the communication line is n times greater than the band width of the modulating function (n is the frequency modulation index). However, after demodulation, the receiver reconstitutes a signal with a band width $\Delta\nu$ corresponding to the band width of the modulating function $x(t)$.

Relation of pulse length to pulse band width. Number of pulses transmitted through a channel of given band width Δf

A basic relation exists between the pulse length and the band width of the pulse spectrum $\Delta\nu$:

$$\tau\,\Delta\nu = \text{const.} \tag{3.10}$$

It follows from this relation that the band width of a pulse is inversely proportional to pulse length. The numerical value of the constant depends on the shape of the pulse. In all cases, however, this constant is of the order of unity, and for some pulses (e.g., square pulses) it may even be taken equal to unity.

Equation (3.10) is a very general relation which is valid for any time-variable process of duration τ. Hence it follows that a continuous time function $x(t)$ with a band width $\Delta\nu$ and duration $\Delta t > \Delta\nu^{-1}$ is of necessity a combination of several individual pulses of various durations $\tau_i < \Delta t$, the shortest of which is of duration τ of the order of $\frac{1}{\Delta\nu}$.

Let us now determine the number of pulses that can be transmitted in unit time through a channel of band width Δf. Let the pulse duration be $\tau_1 = \frac{1}{\Delta f}$. The band width of this pulse is $\Delta\nu_1 = \frac{1}{\tau_1} = \Delta f$ (we took the constant in (3.10) to be equal to 1). Since the channel band width is equal to the pulse band width, all the frequency components of the pulse will be transmitted through the channel and the pulse will be reconstituted without distortion at the receiving end. The total number of pulses transmitted through the channel in unit time is $\frac{1}{\tau_1} = \Delta f$. Now suppose that the pulse is 10 times longer, $\tau_2 = \frac{10}{\Delta f}$. The band width of this pulse is 1/10 of the band width of the previous pulse, $\Delta\nu_2 = \frac{1}{\tau_2} = 0.1\,\Delta f$. Separating the signals in frequency, we can accommodate in our communication line 10 frequency channels of width Δf each. Each of these channels will transmit $\frac{1}{\tau_2} = 0.1\,\Delta f$ pulses in unit time, and the total number of pulses transmitted through all the 10 frequency channels will be Δf as before. Finally, let the pulse duration be $\tau_3 < \frac{1}{\Delta f}$. The band width of each pulse is then greater than Δf. The pulse components with frequencies $\nu > \Delta f$ are not transmitted through the communication channel, and the signal is distorted. It may therefore seem that Δf determines the maximum number of pulses which are transmitted without distortion in 1 sec through the communication channel. However, this is not exactly so; a more rigorous treatment shows that the maximum number of pulses is double this quantity, being equal to $2\Delta f$.

Indeed, let an ideal frequency filter with a pass band Δf be mounted at the entrance to the communication line. At the time $t = 0$, a brief pulse $\left(\tau < \frac{1}{\Delta f}\right)$ of arbitrary shape is delivered to the filter input. After passing through the filter, the pulse becomes blurred and its shape is described by the function

$$x(t) = x_0 \frac{\sin 2\pi \Delta f t}{2\pi \Delta f t}, \qquad (3.11)$$

where x_0 is the amplitude of the original brief pulse. The properties of this function are responsible for the fact that the communication channel is capable of transmitting every second a number of pulses equal to double the channel band width. Function (3.11) is shown graphically in Figure 36. For $t=0$, $x=x_0$; for $t={}^1/_2\Delta f, {}^2/_2\Delta f, {}^3/_2\Delta f, \ldots, x(t)=0$. If we now send a train of brief pulses at equal time intervals $\Delta t = {}^1/_2\Delta f$, we obtain some combination signal, a sum of signals of the form (3.11) displaced by an amount $i\Delta t$ $(i=1, 2, 3\ldots)$ relative to $t=0$. This combination signal has the form

$$\tilde{x}(t) = \sum_i x_i \frac{\sin 2\pi \Delta f \left(t - \frac{i}{2\Delta f}\right)}{2\pi \Delta f \left(t - \frac{i}{2\Delta f}\right)}. \qquad (3.12)$$

Since each term of this sum is equal to zero at any of the times $t_j = j\Delta t$ for $j=1, 2, 3\ldots$ except $j=i$ (see Figure 35), the combination signal at any of the sending times t_i is determined only by the amplitude x_i of the corresponding brief pulse. Thus, despite the distortion of short pulses after transmission through a filter of band width Δf, these pulses following one another at a rate of $2\Delta f$ pulses per second will be fully reconstituted if the pulses at the receiving end of the line are measured at the same rate (at intervals $\Delta t = {}^1/_2\Delta f$).

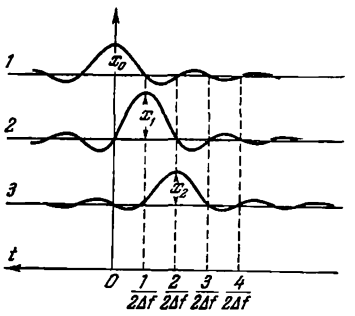


FIGURE 36. Illustrating the determination of the number of pulses transmitted in unit time through a channel of given band width. After transmission of a short pulse of arbitrary shape through an ideal low-frequency filter with a pass band Δf, the pulse is distorted to the shape shown in this figure:

1) $x(t) = x_0 \frac{\sin 2\pi \Delta f t}{2\pi \Delta f t}$

2) $x(t) = x_1 \frac{\sin 2\pi \Delta f \left(t - \frac{1}{2\Delta f}\right)}{2\pi \Delta f \left(t - \frac{1}{2\Delta f}\right)}$

3) $x(t) = x_2 \frac{\sin 2\pi \Delta f \left(t - \frac{2}{2\Delta f}\right)}{2\pi \Delta f \left(t - \frac{2}{2\Delta f}\right)}$

The curves correspond to different pulses with amplitude x_0, x_1, x_2, which are delivered to the filter input at the times $t_0 = 0$, $t_1 = \frac{1}{2\Delta f}$, $t_2 = \frac{2}{2\Delta f}$.

At the time t_i, the amplitude of the i-th pulse (after transmission through the filter) is x_i, and the amplitudes of all the other pulses are zero. The combination signal at the time t_i is therefore entirely determined by the amplitude x_i of the initial signal.

Transmission of continuous functions by pulsed signals

A continuous message function $x(t)$ of duration $\Delta T = t_2 - t_1$ can be represented by a sequence of discrete values $x(t_k)$ taken at time intervals Δt_k. The representation is clearly of higher accuracy for small time intervals Δt_k. The discrete values of the function can be transmitted through the communication channel using one of the pulse modulation systems. If Δf is the channel band width, the maximum number of pulses than can be transmitted in unit time through this channel is $2\Delta f$. Using a succession of pulses following one another at this rate, we obtain at the receiving end a time function $\tilde{x}(t)$ which is expressed by (3.12). In a noise-free channel, the values of this function at the quantization times t_k are determined entirely by the

values of the original function, $\tilde{x}(t_k) = x(t_k)$. The question is, are these functions equal at any time t, and not only at t_k, i.e., are they identically equal? The fit between the two functions is naturally improved if the original function varies slowly between the quantization times t_k. This means that the function should not contain very high harmonics. According to Kotel'nikov's theorem, the two functions are identical if the original function $x(t)$ does not contain components with frequencies ν higher than Δf, i.e., if the band width of the $\Delta\nu$ transmitted function is equal to the band width of the communication channel. Kotel'nikov's theorem is highly significant for the theory and technology of communication, since it permits converting continuous functions into a train of some discrete magnitudes for transmission. This theory maintains that a function with a bounded spectrum $\Delta\nu$ is completely determined by its values measured at intervals $\Delta t = {}^1/_2\Delta\nu$. In particular, a function of duration Δt, i.e., a function which does not vanish only for $t_0 < t < t_0 + \Delta t$, is determined by a set of $2\Delta t\Delta f$ discrete values. Thus, the definition of information derived for discrete messages can be safely applied to continuous functions with a bounded spectrum.

When continuous functions are transmitted by means of pulsed signals, the main difficulty is that the function may take on any instantaneous values, including irrational and transcendental numbers with an infinite number of significant digits. Theoretically (in a noise-free channel), these numbers can be transmitted with full faithfulness by PAM or another suitable technique. In reality, however, reconstitution of the original pulse with sufficient accuracy (or transmission of a sufficiently high number of significant digits) in a noisy channel requires an excessively high signal-to-noise ratio in the communication channel. Therefore, the next step adopted in the transmission of continuous functions calls for quantization of the message. To quantize the message, we select from among all the values of $x(t)$ a set of N discrete allowed levels $x_1, x_2, \ldots, x_N$, which are distant Δx from one another (the quantization gap). All the other values are regarded as forbidden. Only the allowed values are transmitted. If the true instantaneous value of the function falls inside the interval (x_i, x_{i+1}), i.e., takes on a forbidden value, the nearest allowed value, differing from the true value by less than half the quantization gap, is transmitted through the channel. This operation is completely analogous to the rounding-off of numbers; it essentially signifies that we are transmitting the true values of the function up to a certain number of significant digits.

The quantized values of the signal in the communication channel are affected by random noise. The width of the quantization gap should be so chosen that with a given probability p the noise does not exceed half the quantization gap. Then the signal can be accurately reconstituted at the receiving end of the channel, since in this case the signal level nearest to the noise-distorted value is the same as that fed into the communication channel. The probability of signal reconstitution error is equal to the given value p. The reconstituted signal can be again sent through the communication line, and this procedure may be repeated several times, without affecting

the reconstitution of the original quantized level. The transmission of quantized values instead of the true values is equivalent to superimposing a certain noise δ_k which does not exceed half the quantization gap. This noise is known as quantization noise. Quantization thus does not free the

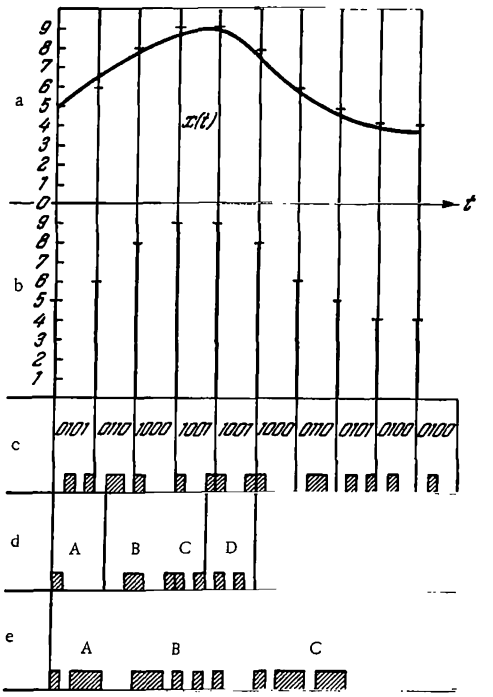


FIGURE 37. Transmission of messages by pulse code:

(a) a continuous message function $x(t)$; (b) quantized values of the function; (c) transmission of the quantized function by binary code; (d) Baudot telegraph code; (e) Morse code.

signal from noise, but in effect substitutes one kind of noise for another. The random uncontrolled noise is replaced with an artificial noise — the quantization noise. The intensity of this noise is not weaker than that of the undesirable natural noise. However, the advantage of a quantized

system is that the noise is fully controllable and the accumulation of random errors is avoided.

Another important advantage of quantization is that it permits transmitting continuous message functions by means of pulse codes. Each discrete value of the function is expressed by a certain positive number, i.e., it can be written in any number system in the form of a certain numerical sequence. The electrical signal corresponding to this discrete value of the function similarly can be represented as a combination of individual electrical pulses. The various pulse combinations corresponding to the various values of the message function constitute a certain code. Every individual combination is regarded as a code combination. Various elementary signals (pulses) used to construct the code combination are known as the code elements, and the number of different elements used in the code combination is the base of the code.

The Morse telegraph code is an example of a ternary code. Its elements are a short signal, "dot", a long signal, "dash", and the absence of a signal, a blank of the same duration as the "dash" intended to separate successive letters. The number of symbols in Morse code combinations is variable; this is a nonuniform code. The Bodo telegraph code is a five-digit binary code; its elements are a pulse and an absence of a pulse, both of equal length. Each message (a letter of the alphabet) is represented by a five-element code combination (Figure 37).

If a is the base of the code, m is the number of elements in a code combination, the total number of code combinations or different values that can be transmitted by this code is $N=a^m$. Quantization makes it possible to transmit functions using a code with a finite number of elements m in each code combination. Without quantization, $N=\infty$ and in general an infinite number of code elements in a code combination will be needed to transmit the true instantaneous values of the function $x(t_k)$.

The transmission of continuous functions by a pulse code whose elements are pulses differing in their height h only is known as pulse-code modulation.* In PCM transmission, the signal is first confined to a limited band width, so that all the frequencies above a certain ν_0 are cut off. Signal readings are then taken at a rate of $2\nu_0$ per second. The readings and quantized and encoded. The number of elements m in a code combination for a given base a is determined by the required number N of quantum levels. Thus, in telephone communications, the best sound is achieved for $N=100$. Therefore, a seven-digit binary code can be used for the PCM transmission of telephone conversation ($2^7 = 128$). Code groups are delivered to the communication line. At the receiving end, the pulses distorted by noise are reconstituted, the code groups are decoded, and a new sequence of pulses with amplitudes proportional to the initial quantum values $x_1, x_2, \ldots, x_N$ is formed. These pulses, coming at a rate of $2\nu_0$ pulses per second, are transmitted through a low-frequency filter with a cutoff frequency ν_0, and are then combined to give the original signal.

* Generally, PCM is regarded as transmission of message by binary pulse code. This technique is often used in practice. However, theoretically, we may consider PCM for any code base.

Let us find the threshold power of PCM. (Threshold power in this case is defined as the threshold mean power, rather than P_{min}.) Consider a code with a general base a; let Δh be the difference in the pulse heights corresponding to two successive elements of this code. The power of the i-th pulse is $P_i = x_i^2 = (i\,\Delta h)^2$, and the mean signal power, assuming a uniform frequency of occurrence of all the pulses, is $P_s = \frac{1}{a}\Sigma P_i$. This power is minimum if both positive and negative pulses are used to make it up. Then,

$$P_s = \frac{(\Delta h)^2}{a} \sum_{i=-\frac{a-1}{2}}^{i=\frac{a-1}{2}} i^2 = \frac{(\Delta h)^2}{12}(a^2-1). \tag{3.13}$$

To ensure correct reconstitution of noise-distorted signals, the random noise ξ should not exceed half the value of $\Delta h\left(|\xi| < \frac{1}{2}\Delta h\right)$. The probability of this even, as we have seen before, depends on the ratio $\frac{\Delta h}{\sigma}$. For $\Delta h = 10\sigma$ the probability of an error (i.e., an incorrect reconstitution of the pulse) is 10^{-6}. Inserting this value of Δh in (3.13), we obtain the threshold power $P_s^0(a)$ for a PCM system with a code of base a:

$$P_s^0(a) = \frac{100}{12}\sigma^2(a^2-1) = \frac{100}{12}(a^2-1)P_n. \tag{3.14}$$

For a given noise power, the threshold signal power increases with the increase of code base. The maximum threshold power is observed for $a=N$, i.e., for ordinary PAM. The threshold power of PAM is

$$P_{PAM} = \frac{100}{12}(N^2-1)P_n \approx N^2\frac{100}{12}P_n. \tag{3.15}$$

The threshold power in this case is seen to be proportional to N^2. For any other code base $a \neq N$, the threshold power is independent of the number of quantization levels N and is determined by the code base only. For a given a, there should be m pulses in each code group to encode the quantum levels (since $N = a^m$). If we reduce the base a, m is increased correspondingly, i.e., the number of pulses transmitted through the line in unit time increases. PCM thus enables us to reduce the threshold signal power by increasing the band width of the communication line. The minimum threshold power is attained when using binary code. In this case $P_s^0(2) = 25P_n$.

Transmission rate of a communication channel

We have now reached the stage when the transmission rate of a communication channel can be determined. On p. 76 we mentioned that the transmission rate is proportional to the channel band width. The signal-to-noise ratio also plays an important part in this respect.

Consider a message which constitutes a table of three-digit decimal numbers. We have a channel of 3 kHz band width and a signal-to-noise ratio $\frac{P_s}{P_n} = 25$. Using (3.14), we find that for this signal-to-noise ratio

the code base is $a=2$, and from the relation $N=10^3=a^m$ we find $m=10$, i. e., a ten digit binary code can be used to transmit the message through our channel. A channel with 3 kHz band width will transmit 6000 pulses per second, or 600 code groups of 10 bits each. The quantity of information contained in each code group is $Q_1=3\log_2 10=10\log_2 2=10$ bits. The transmission rate of the channel is therefore 6000 bits per second. Now suppose that the transmitter power is increased by a factor of 5, so that the signal-to-noise ratio becomes $\frac{P_s}{P_n}=125$. If we are using binary code, as before, the channel transmission rate for the given band width (3 kHz) naturally does not change. However, the binary code is not very efficient for such a high signal-to-noise ratio. The transmission rate can be raised by using a different code system. From (3.14) we find that for $\frac{P_s}{P_n}=125$, we may take $a=4$. Now from $N=10^3=a^m$, we get $m=5$. We may thus use a five-digit quaternary code. Transmitting as before 6000 pulses per second, we may now transmit 1200 code groups of five quaternary pulses each. The quantity of information associated with each code group is 10 bits as before ($3\log_2 10=5\log_2 4\approx 10$) and the transmission rate is therefore $10\times 1200=12{,}000$ bits per second. For $\frac{P_s}{P_n}\approx 825$, we may use a three-digit decimal code, raising the transmission rate to $10\times 2000=$ $=20{,}000$ bits/sec. Finally for a signal-to-noise ratio equal to $8\cdot 10^6$, we may take $a=10^3=N$, $m=1$, i. e., transmit using the ordinary PAM (without coding). Each pulse corresponds to a three-digit decimal number and thus contains 10 bits of information. A channel of 3 kHz band width may transmit 6000 such pulses and the transmission rate of the channel will therefore be $6\cdot 10^4$ bits per second. The same quantity of information can be transmitted for $\frac{P_s}{P_n}=25$, using a binary code and increasing the channel band width from 3 to 30 kHz.

This example clearly illustrates the importance of each factor affecting the channel transmission rate. The frequency band determines the number of pulses that can be transmitted through the channel in unit time. The signal-to-noise ratio gives the base of the code that may be used for transmission through the particular channel and, hence, the information content of each pulse. Thus, in binary code transmission, each pulse carries 1 bit of information, with ternary code each pulse carries 1.6 bits, in quaternary code 2 bits, in decimal code 3.3 bits, etc. By reducing the code base, we lower threshold power of the system and at the same time lower the quantity of information carried by each signal, so that to ensure a constant transmission rate the band width must be increased.

Let us find the transmission rate of a PCM channel. Let the band width of the communication line be Δf. Then it will carry $2\Delta f=nm$ pulses per second, where n is the number of code groups transmitted each second through the communication channel, and m is the number of pulses in the code group. The information Q_1 associated with the transmission of each code group is $Q_1=m\log_2 a$, and the total quantity of information transmitted through the channel in 1 sec is

$$q=nQ_1=nm\log_2 a=2\Delta f\log_2 a=\Delta f\log_2 a^2. \qquad (3.16)$$

Inserting a^2 from (3.14), we obtain

$$q = \Delta f \log_2 \left[1 + \frac{12}{100} \frac{P_s^0(a)}{P_n} \right]. \tag{3.17}$$

This is the maximum transmission rate of a PCM system. If the code base a is chosen so that $P_s^0(a)$ is the mean signal power in the communication line, the $P_s^0(a)$ in (3.17) can be replaced by P_s. For any other code base b $(2 \leqslant b \leqslant a)$,

$$2\Delta f \leqslant q \leqslant \Delta f \log_2 \left(1 + \frac{12}{100} \frac{P_s}{P_n} \right). \tag{3.18}$$

A useful characteristic of a communication system is the ratio $\frac{q}{\Delta f}$, which characterizes the transmission rate per 1 Hz. The corresponding values for PCM are listed in Table 3.1.

TABLE 3.1. PCM transmission rate per unit band width

Code base, a	Threshold power P_s/P_n	Number of bits per pulse	Number of bits per 1 Hz $\frac{q}{\Delta f} = \log_2 \left(1 + \frac{12}{100} \frac{P_s}{P_n} \right)$
2	25	1.0	2.0
3	67	1.6	3.2
4	125	2.0	4.0
5	200	2.3	4.6
6	292	2.6	5.2
7	400	2.8	5.6
8	525	3.0	6.0
9	666	3.2	6.4
10	825	3.3	6.6

The PCM coding system is not optimal. The transmission rate expressed by (3.17) therefore does not realize the full potential of the communication system. Shannon /5/ has shown that there exists some coding system, which in general may be quite complex, for which the transmission rate can be raised to

$$q = \Delta f \log_2 \left(1 + \frac{P_s}{P_n} \right). \tag{3.19}$$

This coding system is termed ideal.

Shannon's equation (3.19) described the maximum transmission rate of a channel of given band width Δf and given signal-to-noise ratio $\frac{P_s}{P_n}$. No communication system, however complex and sophisticated, will transmit information at a higher rate for the same Δf and $\frac{P_s}{P_n}$. Shannon's formula thus establishes the limiting relation between the basic parameters of a communication system, systems of communication with extraterrestrial civilizations included.

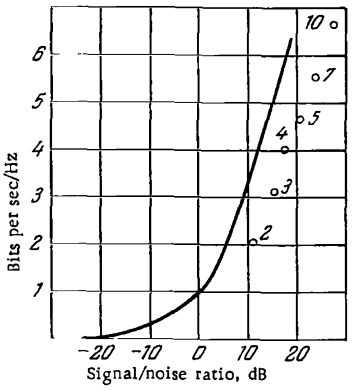


FIGURE 38. The transmission rate of a communication channel.

The solid curve corresponds to Shannon's ideal system. The dots refer to PCM with positive and negative pulses for error frequency of 10^{-5}; the numerals next to the PCM dots correspond to the code base.

Figure 38 plots the rate of information transmission per 1 Hz, $\frac{q}{\Delta f}$, as a function of the signal-to-noise ratio in a communication channel for an ideal Shannon system and for PCM. The ideal coding system ensures a gain of 8 — 10 dB in power compared to the PCM. Moreover, the PCM has a sharp threshold power $P_s^0(a)$, determined by the assumed error probability. Both the threshold power and the associated numerical coefficient before $\frac{P_s}{P_n}$ in (3.17) change when the error probability is changed. For $P_s < P_s^0(a)$, information cannot be transmitted with the specified reliability (the specified error frequency). An ideal system does not have a clearcut threshold power. It may operate for any P_s, ensuring reliable transmission of information according to (3.19) with any arbitrarily small error probability. In particular, for $\frac{P_s}{P_n} = 3$, $q = 2\Delta f$, i.e., the ideal system has a transmission rate equal to the transmission rate of a binary PCM system (for a threshold signal-to-noise ratio $\frac{P_s}{P_n} = 25$). For $\frac{P_s}{P_n} = 1$, $q = \Delta f$ for the ideal Shannon system, and then it rapidly decreases, reaching zero for $\frac{P_s}{P_n} = 0$. Finally, for $P_n \to 0$, $\frac{P_s}{P_n} \to \infty$ and q also goes to infinity, i.e., the rate of information transmission through a noise-free channel can be made arbitrarily large. This is also true for PCM. In practice, this feature can be realized in PCM systems by using a code with a very large base a. Indeed, any text may be represented as a number with sufficiently numerous significant digits. This number can be transmitted through a noise-free channel as a pulse of appropriate height.

Let us consider the dependence of the maximum transmission rate of a channel on band width. The P_n entering Shannon's formula (3.19) depends on the band width. In most practical cases, we may take

$$P_n = P_{n.sp}\Delta f. \tag{3.20}$$

Here $P_{n.sp}$ is the noise power per unit frequency interval, called the specific noise power. Inserting P_n from (3.20) in (3.19), we find

$$q = \Delta f \log_2\left(1 + \frac{P_s}{P_{n.sp}\Delta f}\right). \tag{3.21}$$

If we take $\Delta f_0 = \frac{P_c}{P_{n.sp}}$, i. e., define Δf_0 as the band width for which the noise power is equal to the signal power, we may write (3.21) in the form

$$\frac{q}{\Delta f_0} = \frac{\Delta f}{\Delta f_0} \log_2\left(1 + \frac{\Delta f_0}{\Delta f}\right). \tag{3.22}$$

Figure 39 shows $\frac{q}{\Delta f_0}$ as a function of $\frac{\Delta f}{\Delta f_0}$.

As the band width increases, the transmission rate rapidly grows up to a point where the signal power becomes comparable to the noise power (for $\Delta f = \Delta f_0$). After that point, the growth of the transmission rate is slowed down, and for $\Delta f \to \infty$, it goes asymptotically to the transmission rate for $\Delta f = \Delta f_0$ multiplied by $\log_2 e = 1.443$.

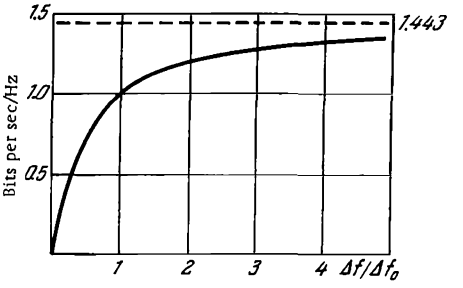


FIGURE 39. The transmission rate of a communication channel as a function of the band width. Δf_0 is the band width for which $P_n = P_s$.

§ 2. RANGE AND INFORMATION CONTENT OF INTERSTELLAR COMMUNICATION

The optimum communication frequencies

We have considered some applications of the general theory of communication, and now we can proceed with a discussion of the problem of communication with extraterrestrial civilizations. The main difficulty of setting up a system of communication with extraterrestrial civilizations is that different elements of the system belong to different "subscribers," and we have no advance knowledge of the type of instruments they are using. As a result, every subscriber, whether on a transmitting or a receiving end, should see to it that the signal transmission and reception devices ensure reliable radio communication despite this intrinsic uncertainty. In this general formulation, the problem includes the various aspects of coding, call signals, signal detection (including the criteria of artificial origin of signals), and signal decoding. Some of these topics are considered elsewhere in the book.

A schematic diagram of an interstellar radio-communication system is shown in Figure 40. The message from the sender (EC-1) is delivered to a transmitter, which converts it into a signal, and the signal is then radiated into the outer space by the transmitting antenna A_1. At the receiving end of the communication line, the radio waves are picked up by a receiver antenna A_2 and the electric signal is directed to the receiver where, after various transformations, the original message is reconstituted.

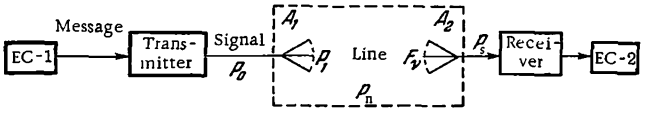


FIGURE 40. A diagram of a system for interstellar radio communication.

A_1 — transmitting antenna, A_2 — receiving antenna, P_0 — transmitter power, P_1 — antenna radiation power, F_ν — special flux density at observation point, P_s — signal power at receiver input.

The communication line is the common element of the system joining the two "subscribers." In interstellar radio communication, the line comprises the part of the outer space between the transmitting and the receiving antennas (the interstellar medium plus the corresponding planetary atmospheres) where the radio waves propagate. The line parameters depend on the conscious activity of the "subscribers," as well as on certain objective factors, such as radio wave absorption in the interstellar medium. We have seen in Chapter II that the absorption coefficient of the interstellar medium increases with the decrease in frequency. Over large distances (of the order of the galactic diameter), the interstellar medium is virtually opaque at meter wavelengths. This automatically limits the range of wavelengths for interstellar communication: because of strong absorption, interstellar communication is unfeasible at frequencies shorter than 1 MHz.

Another important objective factor is the noise in the communication line. The various noises can be divided into two groups: instrumental noise and background noise. Instrumental noise is controllable and it can be reduced to a comfortably low level. Background noise is determined by the radio emission of the planetary atmospheres and the radio waves originating in the outer space. Atmospheric noise in principle can be eliminated by mounting the antennas at an appropriate distance from the planetary surface, e.g., on artificial satellites. Noise associated with radio waves from space is intrinsically unavoidable.

Another source of intrinscially unavoidable noise are the quantum fluctuations,* associated with the quantum nature of the electromagnetic radiation.

* Not to be confused with quantization noise (§1).

Background noise and quantum fluctuations determine the optimum frequency range of electromagnetic waves for interstellar communication. This problem was analyzed in some detail in Chapter I. We have seen that the optimum frequency range for the purposes of the search for call signals of extraterrestrial civilizations is confined to the region of minimum background noise ($\lambda \approx 10-50$ cm), and for reception of meaningful messages to the region of minimum sky brightness temperatures. The last condition is satisfied for a very wide range of frequencies, from decimeter to submillimeter waves.

The choice of the exact working frequency band in the optimum frequency range requires a separate discussion. This topic was also analyzed in Chapter I, where we derived an expression for the optimum distribution of the transmitter energy in the spectrum, needed to ensure maximum information transmission rate. For moderate quantities of information, the question of the transmission rate is not particularly acute, and the frequency band may be taken fairly narrow. In this case, we are faced with the problem of frequency scanning in our search for signals. Cocconi and Morrison /1/ proposed using the frequency of the hydrogen radio line at 21 cm ($\nu = 1420$ MHz) or one of its harmonics. Similarly, the frequency of the hydroxyl OH radio line at 18 cm can be used. Troitskii /6/ suggested that the search should be conducted near the radio lines of individual molecules used in masers (the 1.25 cm ammonia line and the 0.4 cm formaldehyde line).

Range of communication

An important parameter of a communication line is its length or extent. Since to first approximation we may assume that the civilizations are uniformly distributed in space, the number of probable subscribers and, hence, the probability of establishing communication is proportional to the cube of the communication range. What factors determine the communication range? The first step is to define exactly the concept of communication range. We are dealing with two problems: detection of EC signals and reception of meaningful messages. Accordingly, we will discuss the range of detection and the range of communication for the reception of meaningful messages. Before any meaningful information can be received, we have to detect the EC signals. However, the expression for the range of communication is simpler to derive, and we will therefore start with this concept.

The range of communication is equal to the maximum distance over which the communication system is capable of transmitting and receiving information with a given reliability (a given error probability). Over greater distances, the signal power falls below the threshold value, and the signal cannot be reconstituted with the required reliability.

Let us now derive an expression for the communication range. Let P_0 be the power of the EC-1 transmitter, Δf_1 the frequency band of the transmitter, η the efficiency of the transmitting antenna. The power radiated by the antenna is then $P_1 = \eta P_0$. If this power is radiated isotropically,

i. e., uniformly in all directions, the radio flux at the observation point at a distance R is

$$F_\nu \Delta f_1 = \frac{P_1}{4\pi R^2}. \tag{3.23}$$

Here F_ν is the spectral energy flux density, or the energy flux per unit frequency band.

Real antennas are not ideally isotropic: they have certain directional properties. The directional properties of the antenna are characterized by its directivity pattern or diagram. The directivity pattern of the transmitting antenna is a polar diagram which plots the energy flux radiated by the antenna in various directions. Figure 41 shows the directivity pattern of a reflector antenna. Almost the entire energy is radiated by this antenna within a certain small solid angle accommodated by the main lobe of the pattern. If the antenna has a rectangular cross section with sides l and h, the angular width of the main lobe in the corresponding directions is

$$2\theta_l = 2\frac{\lambda}{l} \text{ and } 2\theta_h = 2\frac{\lambda}{h}, \tag{3.24}$$

where λ is the wavelength. For a circular reflector antenna (e.g., a paraboloid of revolution), the width of the main lobe is

$$2\theta_0 = 2 \times 1.22\frac{\lambda}{D} = 1.22\frac{\lambda}{r}, \tag{3.25}$$

where r is the radius, D is the reflector diameter. This quantity is usually referred to as the width of the antenna pattern or the beam width at zero power level. Another significant parameter is the beam width between half-power points, which for a circular cross section antenna is expressed by the equality

$$2\theta_{0.5} = 2 \times 0.51\frac{\lambda}{D} \approx \frac{\lambda}{D}. \tag{3.26}$$

To first approximation, the antenna pattern may be regarded as constant (equal to its maximum value) within the beam width angle, falling to zero outside this angle.

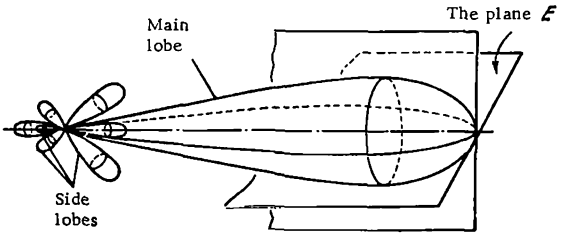


FIGURE 41. The directivity pattern of an antenna.

In calculations of the radiated power, we will use the directivity coefficient of the antenna. The directivity coefficient of a transmitting antenna is equal to the ratio of the antenna power radiated in a certain direction (e.g., along the axis) in a unit solid angle to the mean power radiated in a unit solid angle in all directions. In other words, the directivity coefficient is defined as the ratio of the energy flux radiated by the antenna inside a small angle $d\omega$ to the energy flux radiated by an isotropic radiator of the same power in the same solid angle $d\omega$. When using a directional antenna with a directivity coefficient g_1, the radio flux at the observation point at a distance R will be

$$F_\nu \Delta f_1 = \frac{P_1 g_1}{4\pi R^2} = \frac{P_0 \eta g_1}{4\pi R^2} = \frac{P_0 \varepsilon_1}{4\pi R^2}. \tag{3.27}$$

The quantity $\varepsilon_1 = \eta g_1$ is known as the antenna gain. If the transmitter power P_0 and the antenna gain are known, the radio flux can be determined without difficulty at any observation point. In what follows, we will assume for simplicity $\eta = 1, P_1 = P_0, \varepsilon_1 = g_1$.

Receiving antennas are also directional. In the theory of antennas it is proved that, in virtue of the reciprocity principle, the antenna properties are the same in transmission and reception. In particular, the antenna pattern, the directivity coefficient, and the gain of the receiving antenna are equal to those of the same antenna working as a transmitting antenna (when a transmitter is connected to the antenna terminals).

The power P delivered by the antenna to the receiver is clearly proportional to the radio flux at the reception point. We may therefore write

$$P_s = SF_\nu \Delta f. \tag{3.28}$$

S, expressed in cm^2, is the effective area of the receiving antenna. This quantity is equivalent to the exit aperture of an optical telescope. In particular, for a reflector antenna with $\eta = 1$, the effective area is equal to the geometrical area of the reflector. The effective area and the antenna gain are related by the equality

$$S = \frac{\lambda^2}{4\pi} \varepsilon. \tag{3.29}$$

We can now derive an expression for the range of communication as a function of the parameters of the transmitting and the receiving systems. The signal power P_s at the receiver input substantially depends on the ratio of the transmitter to receiver band width. Two possibilities should be considered here.

a) The receiver band width is greater than the transmitter band width ($\Delta f_2 > \Delta f_1$).

This case is observed, e.g., for the reception of narrow-band monochromatic signals. Using (3.27) and (3.28) and introducing the subscript 1 to identify the parameters of the transmitting system and subscript 2 to identify those of the receiving system, we find

$$P_s = S_2 F_\nu \Delta f_1 = S_2 \frac{P_1 g_1}{4\pi R^2}. \tag{3.30}$$

Note that the result is independent of the receiver band width Δf_2 and, for a given transmitter power P_1, it does not depend on the transmitter band width either. The noise power at the receiver input, as we know, is proportional to the receiver band width:

$$P_n = P_{n\cdot sp}\,\Delta f_2 = kT_n\,\Delta f_2. \tag{3.31}$$

Here k is Boltzmann's constant, equal to $1.38 \cdot 10^{-16}$ erg/deg, T_n is the noise temperature, generally introduced as a parameter of the noise power. It is equal to the temperature of an active load (a resistor) matched to the receiver input which produces the same noise power when connected in place of the antenna. When dealing with background noise, T_n is the equivalent brightness temperature of the noise radiation. In particular, if the background is associated with the thermal radio emission of some space medium, T_n coincides with the temperature of that medium.

The last two expressions give the signal-to-noise ratio at the receiver input:

$$\alpha = \frac{P_s}{P_n} = \frac{P_1 g_1 S_2}{4\pi R^2 k T_n\,\Delta f_2}. \tag{3.22}$$

In § 1 we saw that this ratio describes the reliability of communication. For reliable communication, α moreover should exceed a certain threshold value, which depends on the particular coding system used. In usual communication systems, $\alpha > 1$. Equation (3.32) shows that the reliability of interstellar radio communication is proportional to the transmitter power multiplied by the transmitting antenna gain and the effective area of the receiving antenna and is inversely proportional to the noise temperature, the receiver band width, and the square of the distance between the civilizations.

For given α, the distance R at which the required signal-to-noise ratio is attained can be found from (3.32):

$$R = \left(\frac{P_1 g_1 S_2}{4\pi \alpha k T_n\,\Delta f_2}\right)^{1/2}, \tag{3.33a}$$

or, using (3.29),

$$R = \left(\frac{P_1 S_1 S_2}{\alpha k T_n\,\Delta f_2 \lambda^2}\right)^{1/2}, \tag{3.33b}$$

$$R = \left(\frac{P_1 g_1 g_2 \lambda^2}{16\pi^2 \alpha k T_n\,\Delta f_2}\right)^{1/2}, \tag{3.33c}$$

i.e., the range of radio communication increases with the increase in the transmitter power and the directivity or the effective area of the receiving and the transmitting antennas; it also increases with the decrease in noise temperature and the receiver band width. The dependence on λ in (3.33b) is attributed to the fact that, for a given area S_1 of the transmitting antenna, the directivity increases at shorter wavelengths;

the dependence on λ in (3.33c) is associated with the fact that, for a given g_2, the effective area of the receiving antenna increases with the increase in wavelength.

b) Let us consider the second case: the transmitter band width is greater than the receiver band width ($\Delta f_1 > \Delta f_2$). This case is observed for the reception of wide-band signals, e. g., when the transmitter energy distribution is determined by the requirement of maximum information content (see Chapter I). The spectrum of the signal in this case is limited by the receiver band width, and the receiver P_s is given by

$$P_s = S_2 F_v \Delta f_2 = \frac{P_1 g_1 S_2 \Delta f_2}{4\pi R^2 \Delta f_1}, \tag{3.34}$$

i. e., in distinction from case a, the signal power is proportional to the receiver band width Δf_2, and for a given total transmitter power, it is inversely proportional to the transmitter band width. The noise power is expressed by (3.31), as before, so that the signal-to-noise ratio (for a given range) and the communication range (for a given signal-to-noise ratio) are respectively given by

$$\frac{P_s}{P_n} = \frac{P_1 g_1 S_2}{4\pi R^2 \Delta f_1 k T_n}, \tag{3.35}$$

$$R = \left(\frac{P_1 g_1 S_2}{4\pi\alpha \Delta f_1 k T_n}\right)^{1/2}. \tag{3.36}$$

Comparison of these expressions with (3.32) and (3.33) shows that they differ only in the subscripts of Δf. In the former case, the signal-to-noise ratio and the range of communication increase with decreasing receiver band width and, for a given transmitter power, are independent of the transmitter band width. In the latter case, conversely, the signal-to-noise ratio and the resulting range of communication increase with the decreasing transmitter band width and are independent of the receiver band width. In general, we may thus write

$$\alpha \propto \Delta f^{-1}, \quad R \propto \Delta f^{-1/2}, \tag{3.37}$$

where Δf is the greater of the two band widths Δf_1 and Δf_2.

Let us consider the range of communication as a function of the parameters of the transmitting and the receiving antennas. This dependence is expressed by (3.33), where Δf_2 should be replaced with $\Delta f = \max(\Delta f_1, \Delta f_2)$. Setting $g_1 = g_2 = 1$ in (3.33c), we obtain the range for the case of isotropic transmission and nondirectional reception. Taking $g_2 = 1$, we obtain the range for directional transmission and nondirectional reception. Finally, taking $g_1 = 1$ in (3.33a), we obtain the range for isotropic transmission and reception with a directional antenna of effective area S_2.

Let us consider the dependence of α and R on band width. Equations (3.32) and (3.33) are conveniently written in logarithmic form:

$$\lg R = \frac{1}{2}\lg P_1 g_1 + \frac{1}{2}\lg \frac{S_2}{4\pi\alpha k T_n} - \frac{1}{2}\lg \Delta f. \tag{3.38}$$

The gain (attenuation) is generally expressed in decibels. The product P_1g_1 can be expressed in dB · W. Let $P_1g_1 = 100$ dB · W. This means that a 1 W transmitter is coupled to an antenna with a 100 dB gain, or alternatively a 1 kW transmitter is coupled to an antenna with 70 dB gain, or finally a 1 MW transmitter is coupled to an antenna of 40 dB gain, etc. If the product P_1g_1 in (3.38) is expressed in dB · W, R in light years, S_2 in square meters, the equation takes the form

$$\lg R = \frac{P_1g_1}{20} + \frac{1}{2}\lg\frac{S_2}{4\pi\alpha kT_n} - \frac{1}{2}\lg\Delta f - 16. \tag{3.39}$$

Let $P_1g_1 = 200$ dB · W, $S_2 = 10^4$ m^2, $T_n = 10°$K; then

$$\lg R = 6.4 - \frac{1}{2}\lg\alpha - \frac{1}{2}\lg\Delta f. \tag{3.40}$$

Similarly, for the same parameters of the receiving and transmitting systems, we obtain

$$\lg\alpha = 12.8 - 2\lg R - \lg\Delta f. \tag{3.41}$$

Figure 42 plots the dependence of α and R on band width. For $\Delta f_2 > \Delta f_1$, the receiver band width clearly should be reduced. The noise power at the receiver input will also decrease, so that the effective power will not change. As a result, the signal-to-noise ratio at a given distance R or the range of communication for a given signal-to-noise ratio will increase. This increase does not entail a loss of information content, since the band width of the communication line (the factor determining the channel transmission rate) in this case is limited by the transmitter band width Δf_1. Moreover, at a given distance, the transmission rate may be increased by increasing the signal-to-noise ratio. The maximum range is attained for $\Delta f_2 = \Delta f_1$. Further decrease of the receiver band width is inadvisable, since the noise and the effective signal will then increase to the same extent. Moreover, further decrease of the receiver band width will limit the transmission band width of the communication channel.

For $\Delta f_2 < \Delta f_1$, the signal entering the receiver is limited on the frequency scale and distorted (e.g., in pulse modulation, the chopping of the frequency band will cause blurring and interference of the pulses). Moreover, the narrower band reduces the transmission rate of the communication channel. If the full band width Δf_1 of the line is utilized at the transmitting end, a decrease of the frequency band width will result in a partial loss of information. In general, some information is also lost when the transmitter band width is only partly utilized, since the character of the signal and its time and frequency distributions are not known in advance. Hence it follows that the case $\Delta f_2 < \Delta f_1$ is unfavorable for communication. To avoid signal distortion and loss of information in this case, we should increase the receiver band width Δf_2 to Δf_1. This increase of band width will not affect the range of communication, since it increases both the effective signal and the noise. However, for technical reasons, the receiver band width cannot be increased indefinitely. Even special wide-band receivers can hardly be expected to have a frequency band wider than 10% of the particular electromagnetic frequency used. Another technique of band matching calls for

reducing the transmitter band width. This can be achieved in two ways: a) without altering the transmitter specific power, i. e., the power per unit frequency, and b) without altering the total transmitter power P_1.

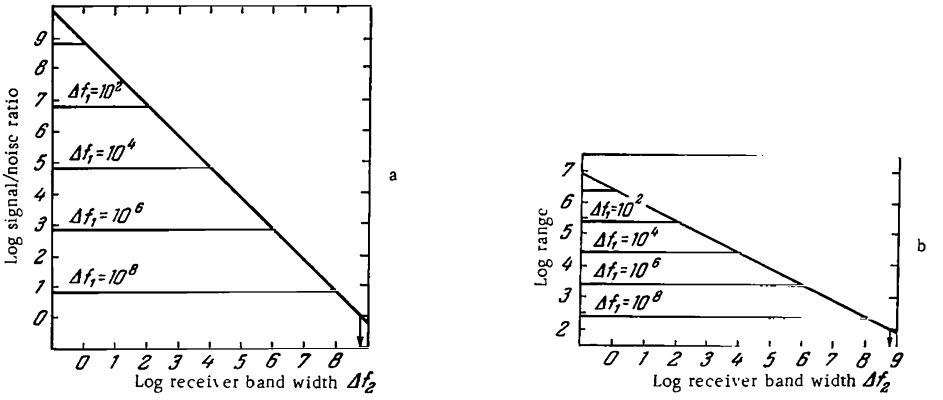


FIGURE 42a. Signal-to-noise ratio α vs. receiver band width Δf_2 for various transmitter band widths Δf_1.

The following system parameters were used: $R = 100$ light years, $P_1 g_1 = 200$ dB·W, $S_2 = 10^4 m^2$, $T_n = 10°K$. The band widths are expressed in Hz. For a given transmitter band width Δf_1, the signal-to-noise ratio increases with the decrease in Δf_2 until the equality $\Delta f_1 = \Delta f_2$ is attained. Further decrease of the receiver band width Δf_2 does not increase the signal-to-noise ratio α. When the band widths are equal, further increase of the signal-to-noise ratio can be attained only by a simultaneous reduction of both the receiver and the transmitter band widths.

FIGURE 42b. Communication range as a function of the receiver band width for various transmitter band widths Δf_1.

$P_1 g_1 = 200$ dB·W, $S_2 = 10^4 m^2$, $T_n = 10°K$. The range is expressed in light years. In both figures the arrow marks the band width for which the signal is equal to noise at a distance of 100 light years (using the given communication parameters).

In case a, the total power and the flux at the observation point decrease in proportion to the decrease in the band width, but the spectral density remains unchanged. The fraction of the total flux or the fraction of the transmitter power delivered to the receiver will thus increase, since the signal-to-noise ratio and the range of communication are not affected. In case b, the contraction of the frequency band entails a growth of the specific transmitter power and the spectral flux density F_ν at the observation point. The total flux $F_\nu \Delta f_1$ remains unchanged, but the fraction of the total flux intercepted by the receiver increases with the decrease in Δf_1. As a result, the contraction of the frequency band will increase the signal-to-noise ratio and the communication range, as we see from (3.35) and (3.36). The maximum range, as before, is attained for $\Delta f_1 = \Delta f_2$. Further contraction of the transmitter band width is inadvisable, since the entire flux at the observation point is anyhow intercepted by the receiver and, for a given transmitter power P_1, the signal-to-noise ratio remains unchanged. If, however, the frequency band is reduced without retaining a constant specific transmitter power, both the signal-to-noise ratio and the communication range will decrease when it falls below Δf_2.

Thus the maximum range of communication is attained for $\Delta f_1 = \Delta f_2$. Once the bands have been made equal (either by reducing the receiver band width for $\Delta f_2 > \Delta f_1$, or by

reducing the transmitter band width for $\Delta f_1 > \Delta f_2$), the communication range can be increased further only by a simultaneous reduction of the transmitter and the receiver band widths. This band width reduction will naturally lead to loss of information. The channel transmission rate for a given separation R between the subscribers will decrease despite the increase in the signal-to-noise ratio, since the dependence of the transmission rate on band width is definitely stronger.

Range of detection

For ordinary systems, the range of communication is limited by the condition $\alpha > 1$.* If, however, we are concerned merely with the detection of signals, without decoding the information that these signals carry, this condition is not necessary. The modern radiometric techniques make it possible to detect signals which are much weaker than noise. This is a common practice in radio astronomy, which deals with extremely weak fluxes from sources in outer space.

The possibility of detecting such signals is based on the statistical properties of noise. Had the noise power been constant, i. e., without any fluctuations in time, it could have been easily corrected by introducing an appropriate voltage in the source, equal in magnitude to the noise voltage and having opposite polarity. In principle, we could thus measure signals of arbitrarily small power level. The measurement procedure reduces to the recording of the small increment above the constant noise level associated with the reception of the effective signal. Correction, strictly speaking, is not absolutely essential: it only constitutes one of the more convenient measurement methods.

Real noise, however, is a random process with voltage (or current) amplitudes fluctuating at random about the zero level. If Δf_2 is the receiver band width, the mean duration of a single noise pulse Δt_2 (or the time during which the amplitude of the damped oscillations generated by this fluctuation remains constant) is of the order of $\frac{1}{\Delta f_2}$. The number of independent noise pulses observed in a time τ_2 is thus $n = \frac{\tau_2}{\Delta t_2} = \tau_2 \Delta f_2$. A recorder with a time constant τ_2 averages these noise pulses, and the mean noise power $P_{n\ av} = \frac{1}{n} \sum_{i=1}^{n} P_i$ fluctuates (so-called recorder fluctuations) about the theoretical mean noise power P_n, which is the expectation value of the random power values P_i of the individual noise pulses. To detect a useful signal, the resulting noise power increment $\Delta P_n = P_s$ should exceed the root-mean-square deviation of $P_{n\ av}$ from the theoretical value P_n, i. e., we should have $P_s = \Delta P_n > \sigma_{av}(P)$ or

$$P_s = \beta \sigma_{av}(P), \tag{3.42}$$

* For Shannon's ideal system, this restriction is of no consequence. Some special communication systems also use methods which permit reception of messages although the signal is much weaker than the noise.

where β is some dimensionless number greater than unity ($\beta > 1$). If we are dealing with Gaussian noise, with a normal amplitude distribution, we may write

$$\frac{\sigma_{av}(P)}{P_n} = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{\tau_2 \Delta f_2}}, \tag{3.43}$$

whence

$$\alpha = \frac{P_s}{P_n} = \beta \frac{\sigma_{av}(P)}{P_n} = \frac{\beta}{\sqrt{\tau_2 \Delta f_2}}. \tag{3.44}$$

This expression determines the minimum signal that can be recorded with a radiometer. For $\beta = 1$, we obtain the limiting or the theoretical radiometer sensitivity. The actual sensitivity is generally much lower, since for reliable signal recording, β should be greater than 10. The factor $\sqrt{\tau_2 \Delta f_2}$ is known as the radiometer gain. For $\sqrt{\tau_2 \Delta f_2} \gg 1$, the signal-to-noise ratio may be much less than unity. For example, for $\tau_2 = 1$ sec and $\Delta f_2 = 10$ MHz, the radiometer gain is 10^4; if $\beta = 10$, we have $\frac{P_s}{P_n} = 10^{-3}$, i.e., the signal power is one thousandth of the noise power, and yet it is 10 times stronger than the rms noise fluctuations and can be reliably detected.

The detectability of signals which are weak compared to noise is associated with the averaging action of the recorder, which averages the individual noise pulses over a period of time equal to its time constant. The effective signal is naturally also averaged, so that the final result is the signal power averaged over the time τ_2. If the characteristic modulation time τ_1 is less than the time constant τ_2 of the recorder, all the measurements related to signal modulation are smoothed out and the information contained in the signal is completely lost. In this case, we can only identify the presence of some effective signal of mean power P_s. It is in this sense that we will interpret the term "range of detection," as distinct from the range of communication.

Reception of information requires that $\tau_1 \geqslant \tau_2$. Using the relation between the time and the band width, we rewrite this inequality in the form*

$$\Delta f_2 \geqslant n \Delta f_1, \tag{3.45}$$

where n is the number of independent noise pulses averaged by the recorder. We have noted before that for purposes of information reception the receiver band width should not be less than the transmitter band.

* Here we take $\tau_1 = \frac{1}{\Delta f_1}$, i.e., use a coding technique which for a given transmitter band width Δf_1 ensures the maximum transmission rate. In general, when the sender does not fully utilize the transmitter frequency band (intentionally lowering the transmission rate of the communication channel, to ensure a higher reliability at a fixed range and a higher range for fixed reliability), the characteristic modulation time τ_1 can be greater than $\frac{1}{\Delta f_1}$, and condition (3.45) is not satisfied.

Condition (3.45) is stronger than that. It shows that in case of averaging, it is no longer sufficient to ensure a receiver band larger than the transmitter band. The receiver band should be greater than the transmitter band multiplied by the square of the radiometric gain.

What is the actual range of detection? Inserting α from (3.44) into (3.33) and (3.36), we obtain the following expressions for the range of detection:

1) $\Delta f_2 > \Delta f_1$,

$$R = \left(\frac{P_1 g_1 S_2}{4\pi\beta k T_n}\right)^{1/2} \left(\frac{\tau_2}{\Delta f_2}\right)^{1/4}, \tag{3.46}$$

2) $\Delta f_2 < \Delta f_1$,

$$R = \left(\frac{P_1 g_1 S_2}{4\pi\beta k T_n \Delta f_1}\right)^{1/2} (\tau_2 \Delta f_2)^{1/4}. \tag{3.47}$$

In the above example, when $\tau_2 = 1$ sec, $\Delta f_2 = 10^8$ Hz, the range of detection can be increased by a factor of 100 due to the radiometric gain. For $\Delta f_2 > \Delta f_1$, the range slowly decreases as the receiver band is made wider; for $\Delta f_2 < \Delta f_1$, it also slowly $(\propto \Delta f_2^{1/4})$ increases, so that it is advisable to increase the receiver band in this case. The maximum range is attained for $\Delta f_2 = \Delta f_1$, as before. In the absence of radiometric gain $(\sqrt{\tau_2 \Delta f_2} = 1)$, equations (3.46) and (3.47) reduce to (3.33) and (3.36), as could have been expected.

Let Δf_2 be the given receiver band width. Consider two signals: a narrow band signal with $\Delta f_{1\,nar} < \Delta f_2$, and a wide band signal with $\Delta f_{1\,wide} > \Delta f_2$. Let R_1 be the range of detection of the narrow-band signal and R_2 the range of detection of the wide-band signal. From the above relations we have $R_1 \propto \Delta f_2^{-1/4}$, $R_2 \propto \Delta f_2^{1/4} \Delta f_1^{-1/2}$ so that

$$\frac{R_1}{R_2} = \left(\frac{\Delta f_{1\,wide}}{\Delta f_2}\right)^{1/2} > 1, \tag{3.48}$$

i. e., for a given receiver band, the range of detection of a narrow-band signal is greater than the range of detection of a wide-band signal. The two ranges R_1 and R_2, however, are not the maximum. In the former case, the range can be increased by reducing the receiver band width to $\Delta f_{1\,nar}$, and in the latter it can be increased by broadening the receiver band width to $\Delta f_{1\,wide}$. We then have

$$R_{1\,max} = R_1 \left(\frac{\Delta f_2}{\Delta f_{1\,nar}}\right)^{1/4}, \quad R_{2\,max} = R_2 \left(\frac{\Delta f_{1\,wide}}{\Delta f_2}\right)^{1/4}, \quad \frac{R_{1\,max}}{R_{2\,max}} = \left(\frac{\Delta f_{1\,wide}}{\Delta f_{1\,nar}}\right)^{1/4}. \tag{3.49}$$

Thus, despite the increase in the radiometric gain with increasing band width, the maximum range of detection of narrow-band signals is greater than the maximum range of detection of the wide-band signals. For example, for $\Delta f_2 = 10^4$ Hz, $\Delta f_{1\,nar} = 1$ Hz, $\Delta f_{1\,wide} = 10^{10}$ Hz, we find $R_1 = 10^3 R_2$; $R_{1\,max} = 10^4\ R_2 = 316\, R_{2\,max}$.

The dependence of the detection range on the band width in a system with averaging is shown in Figure 43. Here, as before, we took $\tau_1 = \frac{1}{\Delta f_1}$, i.e., the sender strives to attain the maximum transmission rate for a given transmitter frequency band. The averaging action gives a gain in range if $\Delta f_2 > \frac{1}{\tau_2}$. The maximum range is attained for $\Delta f_2 = \Delta f_1$. These band widths fall to the right of the line $\Delta f_2 = \frac{1}{\tau_2}$, i.e., in the region where the radiometric gain is greater than unity only for $\tau_1 < \tau_2$, when loss of information occurs. For $\tau_1 > \tau_2$, the band widths corresponding to the maximum detection range fall in the region without radiometric gain. Thus, averaging produces a gain in range while resulting in a complete loss of the information content.

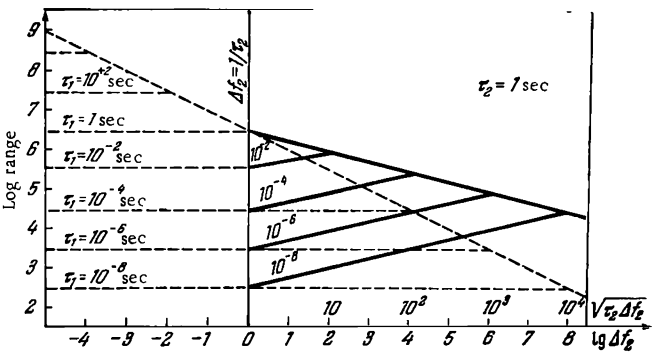


FIGURE 43. Range vs. receiver band width in a system with averaging. The range R is expressed in light years, $P_1 g_1 = 200$ dB · W, $S_2 = 10^4$ m^2, $T_n = 10°$K, $\beta = 1$ $\tau_2 = 1$ sec. For $\tau_1 < \tau_2$ the range of detection at first increases with the decrease of the receiver band width, as long as $\Delta f_2 > \Delta f_1$, and then, passing through a maximum for $\Delta f_2 = \Delta f_1$, starts decreasing: this decrease stops for $\Delta f_2 = \frac{1}{\tau_2}$, when there is no radiometric gain (this is also the situation for $\Delta f_2 < \Delta f_1$ without averaging). For $\tau_1 \geqslant \tau_2$, the range of communication increases with a decrease in receiver band width in proportion to $\Delta f_2^{-1/4}$, up to $\Delta f_2 = \frac{1}{\tau_2}$. If this Δf_2 is still greater than Δf_1, further decrease of the receiver band width is accompanied by a more rapid growth of range (in proportion to $\Delta f_2^{-1/2}$), which stops for $\Delta f_2 = \Delta f_1$. The variation of range without averaging is marked by the dashed lines in the figure.

This remark is applicable to the maximum range. It must be taken into consideration in designing optimum communication channels, when the receiver and transmitter frequencies may be taken equal. The situation is different in communication with extraterrestrial civilizations: we cannot choose $\Delta f_2 = \Delta f_1$, since the transmitter band width is not known in advance. Therefore, if the problem is not confined to the detection of EC signals, but also includes the reception of the information contained in the signals, the receiver band width should be chosen so that it is

not less than the expected transmitter band width, and the time constant should be taken sufficiently large to ensure an adequate radiometric gain. However, τ_2 must nevertheless be smaller than τ_1, i.e., we should have $\frac{1}{\Delta f_2} < \tau_2 < \frac{1}{\Delta f_1}$. Suppose there are grounds to believe that the transmitter band width Δf_1 is of the order of 0.1 Hz. Not to lose any information, we choose the receiver band width with an adequate safety margin, $\Delta f_2 = 1$ Hz. Then, to ensure an averaging gain, the time constant τ_2 should be over 1 sec, but not greater than 10 sec. Let us take $\tau_2 = 4$ sec, and the radiometric gain will then be 2. We have thus obtained a slight gain, not in the detection range, but actually in the communication range, i.e., the range of information reception.*

The question of the assumed transmitter band width is highly uncertain. We are never outside the domain of hypotheses on this topic. The band width may be estimated from considerations regarding the most likely rate of information transmission. If the rate of information transmission is sufficiently low, the working frequency band is limited from below only by the stability of the transmitted signal. In this case, the band may reach a few Hz or fractions of Hz, and in molecular masers even hundredths of Hz.

Range of reception of pulse signals

One of the ways for increasing the range of reception is through the use of pulse signals. If the pulses are widely spaced, a sufficiently high pulse power can be attained with a fairly low-power transmitter. Let Δt_1 be the pulse duration, and t_1 the time between two successive pulses. The ratio of the instantaneous, or so-called peak, pulse power to the mean transmitter power is $\frac{t_1}{\Delta t_1}$. To avoid averaging of the pulses by the recorder, the time constant τ_2 should not exceed the duration of the pulse. If this condition is satisfied, the signal power P_s is proportional to the power pulse. Each sending of length Δt_1 may constitute a simple or a complex pulse. In the case of simple or so-called video pulses (without high-frequency filling), the pulse duration Δt_1 determines the transmitter band width $\Delta f_1 = \frac{1}{\Delta t_1}$. In this case, the condition $\tau_2 \leqslant \Delta t_1$ coincides with (3.45). If $\tau_2 \geqslant t_1$, the signal power P_s is proportional to the mean transmitter power. Therefore, in the range equations we should take

$$P_1 = \begin{cases} P_1, & \text{if} \quad \tau_2 \geqslant t_1, \\ \frac{P_1 t_1}{\Delta t_1}, & \text{if} \quad \tau_2 \leqslant \Delta t_1. \end{cases} \tag{3.50}$$

* If $\tau_1 \gg \frac{1}{\Delta f_1}$ (i.e., when the sender radically reduces the quantity of information sent through the channel in unit time), the receiver band width Δf_2 and the time constant τ_2 can be chosen so that reception of information is ensured for a sufficiently large radiometer gain of the order of $\sqrt{\tau_1 \Delta f_1}$, i.e., in this case, no information is lost in averaging.

The maximum range of communication is attained for $\Delta f_1 = \Delta f_2$, and since in this case, as we have seen, $\tau_2 \Delta f_2 = 1$ (for $\tau_2 \leqslant \Delta t_1$), we obtain

$$R_{\max} = \left(\frac{P_1 g_1 t_1 S_2}{4\pi\alpha \, \Delta t_1 \, \Delta f_1 k \bar{T}_n} \right)^{1/3}. \tag{3.51}$$

In particular, for simple pulses, when $\Delta t_1 \Delta f_1 = 1$,

$$R_{\max} = \left(\frac{P_1 g_1 t_1 S_2}{4\pi\alpha k \bar{T}_n} \right)^{1/3}, \tag{3.52}$$

i. e., the maximum range of communication using simple pulses and a fixed mean transmitter power is independent of the transmitter band width and increases with the increase in the time spacing between the pulses. The feasibility of high-range communication with a relatively low-power transmitter* using widely spaced pulse signals makes this communication technique particularly attractive for interstellar communication. Although the wide pulse spacing lowers the transmission rate of the system, the loss of information is not particularly significant for the transmission of call signals by extraterrestrial civilizations.

Length of transmission. Directivity and information content

As the directivity of the receiving and the transmitting antennas is improved, the signal-to-noise ratio and the resulting range of communication both increase. Should we thus always strive to increase the directivity of the transmitting antennas to the maximum?

Let us consider the relationship between the length of transmission and the directional properties of the antenna, when the exact position of the subscriber is not known in advance. This situation is a good approximation of what we are likely to encounter in communication with extraterrestrial civilizations. Suppose the transmitting antenna is mounted on a planet which, like the Earth, spins with a rotation period T_p and let the antenna axis remain fixed in the planetary system of axes. As a result of rotation, the directional radio beam radiated by the antenna intercepts any given part of the sky for a limited length of time Δt, as long as the corresponding sky area falls inside the main lobe of the antenna pattern. The faster the planetary spin and the higher the antenna directivity, the shorter is the time Δt. Suppose that at some time t the antenna is aimed exactly at the subscriber. In a time $\frac{\Delta t}{2}$, it rotates through the angle

$$\frac{\theta}{2} = \frac{d\theta}{dt} \frac{\Delta t}{2}, \tag{3.53}$$

where $\frac{d\theta}{dt} = \omega_p \cos\delta$ is the velocity of rotation of the beam, ω_p is the angular velocity of rotation of the planet, δ is the angle between the antenna axis and the plane of the equator. Signals are received at the relevant time

* For numerical estimates see Tables 3.3—3.5 and Figure 45.

$t + \frac{\Delta t}{2}$ if the angle of rotation $\frac{\theta}{2}$ does not exceed half the beam width. To fix our ideas, we may set the maximum angle of rotation for which the signals are just received equal to the beam half-width between points of half power $\left(\frac{\theta}{2} = \theta_{0.5}\right)$. Then the total transmission time Δt (from $t - \frac{\Delta t}{2}$ to $t + \frac{\Delta t}{2}$) is

$$\Delta t = \frac{2\theta_{0.5}}{\omega_p \cos \delta}. \tag{3.54}$$

Consider a pencil-beam reflector antenna. The dependence of the beam width on the directivity coefficient for this antenna is expressed by the relation

$$\theta^2 g = \text{const}. \tag{3.55}$$

The numerical value of the constant depends on the exact power level used in reckoning the angle θ. If $\theta = 2\theta_{0.5}$, then const = 10.2; if $\theta = 2\theta_0$, i. e., the total width of the main lobe (or the beam width between zero power points), we have const = 59.2. Using this dependence, we can establish a relationship between the length of transmission Δt and the directivity coefficient of the antenna. Figure 44, borrowed from Webb /7/, gives some idea of the value of Δt for various g in the case of a planet with a 24 hr period of axial rotation, when the antenna axis is aligned in the equatorial plane.

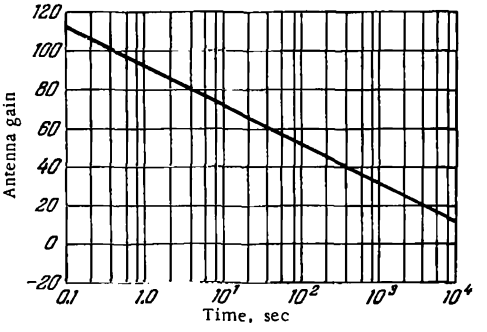


FIGURE 44. Directivity of transmitting antenna as a function of the length of transmission.

The antenna is fixed in the planetary system of axes. The antenna axis is aligned in the equatorial plane, and the planetary rotation period is 24 hrs.

Suppose that after each complete rotation of the planet the antenna axis is displaced in declination through an angle θ equal to the beam width, so

that ever new sky areas are illuminated after each rotation. The total time to scan the entire sky is

$$\Delta T = T_p \frac{\pi}{\theta} \approx T_p \sqrt{g}, \tag{3.56}$$

and the length of transmission received by each subscriber is expressed in terms of the total scanning time in the form

$$\Delta t = \frac{\Delta T}{2g \cos \delta}. \tag{3.57}$$

The minimum duration $\Delta t = \frac{\Delta T}{2g}$ is observed for subscribers located in the equatorial plane. A similar relationship between the total scanning time and the length of transmission to each subscriber is obtained for cases when the scanning is done by moving the antenna proper, without resorting to the planetary rotation; in this sytem, the antenna tracks for a time Δt one given sky area, and is then abruptly aimed at the next area.

Let us now consider the relationship between directivity and information content. Let the transmitter power P_t and the length of the transmission ΔT be given; let P_s be the power signal at the reception point at a given distance R in the case of isotropic transmission. The length of transmission for each subscriber in isotropic sending is equal to the total length of transmission. Therefore, the maximum quantity of information Q_1 that can be transmitted in this time by an isotropic transmitter is given by Shannon's theorem:

$$Q_1 = \Delta f \, \Delta T \log_2 \left(1 + \frac{P_s}{P_n}\right). \tag{3.58}$$

Let us now consider directional transmission with the same P_t and ΔT. The signal power is increased by a factor g due to directivity, and the time of transmission toward each subscriber decreases by the same factor. Therefore, the maximum quantity of information that can be transmitted in a time ΔT by a directional antenna, when the subscriber's position is not known in advance, is given by

$$Q_2 = \Delta f \, \Delta t \log_2 \left(1 + \frac{P_s}{P_n}\right) = \Delta f \frac{\Delta T}{g} \log_2 \left(1 + \frac{gP_s}{P_n}\right). \tag{3.59}$$

Comparison with (3.58) shows that

$$Q_2 \leqslant Q_1. \tag{3.60}$$

The equality is observed when

$$\frac{P_s}{P_n} < \frac{gP_s}{P_n} \ll 1. \tag{3.61}$$

In this case, changing over from binary to natural logarithms and series expanding, we find

$$Q_2 = \frac{gP_s \, \Delta T \, \Delta f}{\ln 2 g P_n} = \frac{P_s \, \Delta T}{\ln 2 P_{n.sp}} = Q_1, \tag{3.62}$$

i. e., for $\frac{P_s}{P_n} \ll 1$, the quantity of information grows in proportion to the length of transmission and the ratio of the signal power to the specific noise power; it is independent of the channel band width.

Equation (3.60) shows that if the required direction of transmission is not known beforehand, the quantity of information decreases on passing from isotropic transmission to directional transmission, all other conditions remaining constant. This conclusion was derived by Siforov /8/. The decrease of information content is associated with the uncertainty in direction to the subscriber. If this uncertainty is reduced, the situation changes radically. Suppose that various considerations (astronomical or other data) indicate that the sending EC should be sought in certain directions in space only. Let $\frac{1}{\gamma}$ be the ratio of the total solid angle Ω corresponding to these "civilized" directions to the entire solid angle 4π. Directional antennas will be more advantageous if

$$\frac{\gamma}{g}\log_2\left(1+\frac{gP_s}{P_n}\right)>\log_2\left(1+\frac{P_s}{P_n}\right). \tag{3.63}$$

In particular, for $P_s=P_n$ and $g\gg 1$, we have

$$\gamma>\frac{0.3g}{\lg g},\quad \frac{\Omega}{(2\theta_{0.5})^2}<4\quad \lg g. \tag{3.64}$$

For $g=10^6$, the size of the region to be scanned with directional transmitting antennas should not exceed the beam area (between points of half power) by more than a factor of 24. For weak signals, the uncertainty in direction may be increased. Let $\alpha=\frac{P_s}{P_n}\ll$ but $\alpha g\gtrsim 1$. Then,

$$\gamma>\frac{0.43\alpha g}{\lg \alpha g},\quad \frac{\Omega}{(2\theta_{0.5})^2}<\frac{3\lg \alpha g}{\alpha}. \tag{3.65}$$

For $g=10^6$ and $\alpha=10^{-2}$, the region of uncertainty (the search region) may reach 1000 times the beam area.

§ 3. CALL SIGNALS AND ARTIFICIALITY CRITERIA

Before establishing communication with extraterrestrial civilizations, we should first detect the sources of artificial signals in outer space. The main difficulty is not that these signals must be picked up against the background of cosmic radio noise (a similar situation is observed in ordinary radio and radar systems): it is that the sources of these signals must be reliably identified and distinguished from a tremendous number of natural radio sources, such as galaxies, radio galaxies, quasars, ionized and neutral hydrogen clouds, supernova remnants, and even individual stars.

To isolate the meaningful radio signals from the jumble of radiation at the receiver input, the incoming radiation must be appropriately

processed. This processing depends on the method of modulation employed by our counterparts. Modern technology provides us with a wealth of means for the analysis of radio waves, but there is no point in applying these analytical tools to sources whose natural origin does not raise any doubts. Before proceeding with the actual analysis, we have to establish that we are dealing with an artificial radio source, or at least there is enough evidence to suspect a source of artificial origin. It would therefore seem that the radiation from an artificial source would possess some peculiar features intended to simplify its detection and identification by other subscribers. Hence the need for a sort of call signal from extraterrestrial civilizations.

We can advance a number of assumptions regarding the likely composition of EC call signals. First, they should ensure a high detection reliability. This condition is best achieved with the aid of continuous (although possibly variable) radio transmission. If the subscriber's position is not known in advance, the transmission should be isotropic, since a highly directional transmitting antenna scanning the sky produces a very short transmission in every given direction (see Figure 44). It moreover seems likely that the call signals contain some information regarding the artificial character of the source, indications of frequency and band width of the transmission, and some additional information which may be regarded as a "key" to the main program. The overall quantity of this information is not particularly large. Therefore, narrow-band quasimonochromatic signals will do as call signals. This is a highly advantageous turn of events, since, on the one hand, a long range of communication is ensured and, on the other, the artificial source can be identified with fair certainty. Indeed, the great majority of the natural radio sources show a very wide, almost unbounded, continuous spectrum. Even the monochromatic radiation of interstellar hydrogen at 21 cm fills a fairly wide band of the order of $5 \cdot 10^4$ Hz. The narrower band of the 18 cm hydroxyl emission, which is assigned to a natural maser mechanism /9/, is a few hundreds of Hz wide. These narrowest natural band widths are clearly inferior to artificial signal generators, which provide band widths of a few Hz or even fractions of a Hz; molecular masers emit in band widths of a few hundredths of Hz. The very detection of such narrow-band signals in itself would provide an indication of a possible artificial origin of the source. Note, however, that the use of narrow-band signals leads to certain difficulties associated with frequency scanning. This problem, however, is not insurmountable, and it will be discussed in the next section.

Along with the narrow-band quasimonochromatic signals, we can expect call signals in the form of widely spaced pulses. This approach also ensures a long range of communication and clearly labels the signal as artificial: natural radio sources generally emit continuously.* Special equipment is required for the detection of these signals.

Although the application of special (narrow-band, pulse, etc.) signals as EC call signals seems to provide the most logical and likely approach to the problem, we cannot rule out another possibility, namely that the transmission will be continuous in a wide frequency band (to ensure a high rate of information transmission), and the function of call

* A remarkable exception to this rule are the pulsars.

signals will be fulfilled by the properties of the source itself and special features of the continuous transmission.

We thus have to solve the problem of the criteria of artificial origin of radio sources. This topic was first attacked by Kardashev /10/. Later it was analyzed by Slysh /11/, Gudzenko and Panovkin /12/, and others. The proposed criteria can be divided into two groups:

1) criteria or signs following from the artificial origin of the source;

2) special properties of radiation, intentionally imposed by the sending EC to ensure communication and simplify detection.

The first group includes such features as angular dimensions, spectrum, statistical properties of signal, variation associated with possible rotation of the system. The second group includes circular polarization, variation associated with modulation, information regarding artificial origin and "keys."

The angular dimensions are one of the most promising and indicative criteria of the first group. The angular size of artificial radio sources cannot exceed a certain (fairly small) value. On the one hand, this is related to the limited scale of activity of civilizations in space (e.g., the scale of a planetary system) and, on the other hand, to the finite speed of propagation of information. Indeed, let t be the time between two successive pulses. To ensure simultaneous emission from different parts of the transmitting system, the distance between the different parts and hence the linear size L of the entire system should not be greater than ct, where c is the velocity of light. If R is the distance to the transmitting system, its apparent angular dimension is

$$\varphi < \frac{ct}{R} = \frac{c}{Rq}, \tag{3.66}$$

where q is the rate of information transmission. For a distance of 1 kpc and $q = 3 \cdot 10^{-4}$ (which corresponds to a transmission of one bit of information per hour), we find $\varphi < 0''.007$. As the rate of information transmission increases, the maximum angular dimension of the source correspondingly decreases. For a rate of 1 bit/sec, $\varphi < 0''.000002$. The angular dimensions of natural sources are generally much larger. Even the less extended sources (the source of the OH line) have angular dimensions of the order of a few thousandths of an angular second. When the steady increase in the sensitivity and the resolving power of radio telescopes will enable us to pick up radio waves from individual stars, this criterion will of course lose some of its paramount importance, but in combination with other signals (power, band width, etc.) it will probably retain much of its value.

If the EC transmitter sends in a sufficiently wide frequency band, its radiation will not be unlike the continuous emission of an artificial source. However, the spectral power distribution of the transmitter will probably differ from the power distribution in the spectrum of natural radio sources. This topic was treated in detail in Chapter I. If the aim is to ensure a maximum transmission rate, the spectrum of the artificial source should look like the curves in Figures 21 and 22. A curve of this shape may be accepted as one of the criteria of artificial origin. This criterion, however, is not very decisive. First, the condition of maximum transmission rate is not absolutely binding. Moreover, excessive saturation of the signal with meaningful information is undesirable, as it interferes with decoding. Second, a similar power spectrum curve may be observed

in some cases for natural sources also. All this notwithstanding, this criterion has its value. In combination with other properties of the radio waves, it may prove to be very useful in establishing the exact nature of the source.

The same considerations apply to signal variation associated with possible rotation of the system. In this case, the length of transmission is determined by the period of rotation and the directivity of the transmitting antenna; the total period of power variation, however, is entirely determined by the rotation period. Variations with periods from a few hours to several days can be expected for transmitters mounted on a spinning planet, and variations with periods from a few months to a few years should be observed for planets or other celestrial bodies which do not spin and only travel around their primary, at a certain distance from it (in the corresponding "zone of life"). For a long time, the opinion prevailed that natural radio sources have a high degree of power constancy. This conclusion emerged from theoretical calculations and there was ample observational evidence to support it. However, after the discovery of the variable radio source CTA-102 /13/, the situation changed radically, since this discovery was soon followed by the detection of the variable radio emission of quasars at various frequencies and with various characteristic times (from a few days to several years). This criterion also has lost its paramount importance, but like the other criteria it should be kept in mind.

The strongest criterion of the first group is apparently that associated with the statistical properties of radiation. This topic was considered by Golei /14/, Slysh /11/, Gudzenko and Panovkin /12/, and Siforov /8/. The radio emission of natural sources is a random, uncorrelated noise, since it is made up of a multitude of independent elementary emission events. In artificial signal generators, on the other hand, the individual emission events are not entirely independent. Therefore, the statistical properties of artificial radiation (e.g., the amplitude distribution) are different from those of noise. The search for artificial radio sources should therefore provide for a comprehensive analysis of the statistical properties of signals.* Analysis of this kind for very weak radio sources is a formidable undertaking. It requires special equipment, different from the conventional tools of the radio astronomer. Note that although the need for a greater emphasis on the statistical analysis of signals has been stressed, little has been done in this direction.

Let us now consider the criteria of the second group. Plane-polarized radiation propagating in the interstellar medium may experience a pronounced rotation of its plane of polarization in the interstellar magnetic fields as a result of the Faraday effect. This is a common phenomenon in radio astronomy, and it is often applied to estimate the distance of the radio source from the observed rotation of the plane of polarization. Although in radio astronomy this is a useful effect, providing additional information regarding the radio source, it is highly harmful in connection with the problem of EC communication, as it definitely distorts the incoming information. The Faraday effect is a sensitive function of frequency.

* The statistical analysis can be based on the moments of the distribution function, the autocorrelation function, the spectral correlation function, etc. /11/.

Therefore, different spectral components of a wide-band signal undergo a different rotation in the interstellar magnetic fields. As a result, the antenna, responding to one direction of polarization only, will record the different spectral components of the signal with different attenuations. The spectrum will be distorted, and the true time characteristics of the signal will become unrecoverable. To avoid this unpleasant effect, the radio transmission sent by the EC should be circularly polarized to start with, and it should be received by a circularly polarized antenna. This is indeed the practice in long-range space communication systems in the solar system.

Variation due to modulation is the most reliable sign of artificial origin of radio signals. The main difficulty, however, is that the characteristic time of the probable power variation is unknown. If the modulation is associated directly with information encoding, the modulation time is probably very short. In binary transmission, with transmission rates of 1000 bit/sec (this is hardly a high rate of transmission: television requires a thousand times higher rate), the characteristic time of power variation, which coincides in this case with the duration of the binary pulse, is 10^{-3} sec. To record such fast variations, we need special equipment with a very small time constant $\tau_2 \leqslant 10^{-3}$ sec. To ensure high sensitivity despite the small time constant, we have to use antennas with a very large effective surface.

The rate of information transmission of call signals may be much lower. Rates of 1 bit/sec are probably more than enough in order to transmit the few tens or hundreds of bits of information probably contained in call signals within a reasonable time. As regards the exact nature of this information, intended to announce the artificial origin of the radio source, we can only guess. Some suggest that several natural or primary numbers can be transmitted to this end; others prefer combinations of known mathematical constants, such as e and π.

Note that special monochromatic signals are not the only candidate for call signals: wide-band signals generated by modulation of short information-carrying pulses will also do. The signal variation, in this case, may correspond to interruptions in the main program, e.g., the beginning or the end of a certain transmission session. These slow power variations of wide-band radio signals can be detected with the existing radio-astronomical equipment. However, it is very important to know the expected period of variation: is it seconds, minutes, or years? This question cannot be answered at this stage. We can only fix a rough lower limit for the probable characteristic time of power variation in the EC call signals. If the transmission is conducted at a frequency ν, the modulation time τ in the EC call signals should satisfy the inequalities

$$\tau > \tau_1 = \Delta\nu^{-1} > \nu^{-1}. \tag{3.67}$$

For radio frequencies, this gives $\tau > 10^{-11} - 10^{-9}$ sec. A more exact estimate can be obtained from the requirement of pulse stability during propagation in the interstellar medium. We have seen in Chapter II that the group delay effect associated with differences in the group velocity for various quasimonochromatic wave groups making up the wide-band pulse imposes certain restrictions on the pulse duration τ. Thus, for a galactic source operating in the range of decimeter wavelength, the pulse

duration should be much greater than 10^{-6} sec (if the source lies outside the plane of the Galaxy) and much greater than 10^{-5} sec (if the source lies in the galactic plane). For an extragalactic source, the limiting pulse duration may reach 10^{-4} sec. Anyhow, we may write

$$\tau > 10^{-6} \text{ sec.} \tag{3.68}$$

The problem of detecting EC call signals would be essentially simplified if we could fix a standard modulation period likely to be used by all EC. This period should naturally satisfy conditions (3.67) and (3.68). We can try to approach this problem by choosing an appropriate combination of universal constants which has the dimension of time or taking as our basis the characteristic time of some processes which are common for the entire Universe, e.g., atomic or comological processes. One of such possibilities is the atomic unit of time equal to the period of orbital revolution of the electron in Born's first orbit, or the so-called Jordan elementary time, equal to the classical radius of the electron divided by the velocity of light. The former quantity is equal to $2.4 \cdot 10^{-17}$ sec, and the latter to $9.4 \cdot 10^{-24}$ sec. However, none of these times satisfies (3.67) nor (3.68). These units of time fix the time scale of microcosmic phenomena. They can be called microscopic time units. On the other hand, there is a completely different megascopic time scale associated with the expansion of the Universe, the time scale characterized by Hubble's constant H, the universal megascopic constant. It would seem that the modulation period in EC call signals should logically fall "half-way" between the microscopic and the megascopic time units; for example, it may be chosen as the geometrical mean of the corresponding numerical values. Taking the same atomic and Jordan elementary time and using the megascopic unit $H^{-1} = 3 \cdot 10^{17}$ sec, we obtain two macroscopic time units

$$\tau_a = \sqrt{2.4 \cdot 10^{-17} \times 3 \cdot 10^{17}} = 3 \text{ sec}$$

and

$$\tau_e = \sqrt{9.4 \cdot 10^{-24} \times 3 \cdot 10^{17}} = 0.002 \text{ sec.}$$

Both these values satisfy conditions (3.67) and (3.68). A logical microscopic time unit is ν^{-1}, where ν is the frequency of the transmitted signal. For the optimum frequency range of interstellar communication $\nu^{-1} = 10^{11} - 10^{-9}$ sec, and the corresponding macroscopic times fall between 30 min and 4.5 hours. The above examples are clearly very sketchy. In particular, the difficulties associated with exact determination of Hubble's constant make it highly unsuitable for use as a basic time unit. It would appear, however, that extraterrestrial civilizations contemplating interstellar communication should have a sufficiently accurate knowledge of it.

We should probably start looking for variations with a period of a few hours. These variations are fairly easy to detect, since no special equipment is needed. These long-term power variations are not distorted by shimmering effects which accompany the propagation of electromagnetic waves in the interstellar and interplanetary medium and in the ionosphere.

The time scale associated with these variations is fairly characteristic of the macrocosmos to which our partners apparently belong (at least if we are dealing with anthropomorphic civilizations). Finally, the very discovery of periodic power variations of period τ related to the radiation frequency by the equality $\tau=\sqrt{\nu^{-1}H^{-1}}$ would attract enormous attention to the corresponding effect.

In conclusion of this section, we would like to stress that the entire topic of call signals and artificiality criteria has hardly been studied so far. Much that is unclear and uncertain remains in this field, opening wide horizons for future research. Rigorous and single-valued criteria should be developed for identifying artificial sources. Such criteria can be based, e. g., on the analysis of the statistical properties of signals or on general theorems of information theory and the theory of complex systems. Some guidelines toward the solution of this problem are indicated in Chapter VI.

§4. METHODS OF DETECTION OF EC SIGNALS

Transmitter power. The power potential of a civilization

In our search for EC signals, we are faced with a two-fold uncertainty: we do not know at what frequency and in what direction these signals are to be sought. A similar uncertainty is in force for the sending EC. The simplest solution to this difficulty is to set up continuous transmission of sufficiently wide-band signals in all directions in space. This ensures simultaneous "service" to all the civilizations within the sphere of action of the transmitter and enables new subscribers to tune in as soon as they reach a suitable technological level. If the signals are sufficiently powerful, and the receiver has a sufficiently high sensitivity, the signals can be received with low-directivity or even isotropic antennas. This has considerable advantages, as it eliminates the need of direction scanning in the first stages of detection. However, this "simple" communication system requires tremendous power. Table 3.2 lists the minimum transmitter power needed for detection and communication using continuous isotropic transmission and undirectional reception at 3 cm wavelength. For detection purposes, the effective signal is assumed to exceed by a factor of 10 the rms noise fluctuation ($\beta=10$), and for the purposes of information reception, the signal is supposed to exceed the noise level by a factor of 100 ($\alpha=100$); the noise temperature was taken equal to 10°K, and the time constant (for detection) $\tau_2=100$ sec. Finally, in accordance with the conditions of minimum power, we took $\Delta f_1=\Delta f_2=\Delta f$. The band width Δf is expressed in Hz, the distance R in light years, the transmitter power P_1 in watts. As we see from the figures in Table 3.2, the required power not only falls far beyond the possibilities of the current transmitters, but actually exceeds the total power potential of mankind.

Mankind is currently consuming annually about $1.5\cdot10^{27}$ erg of energy of various forms, which corresponds to a power of about $5\cdot10^{12}$ watt.

TABLE 3.2. Minimum transmitter power for continuous isotropic transmission and nondirectional reception at 3 cm wavelength ($g_1 = g_2 = 1$; $\Delta f_1 = \Delta f_2 = \Delta f$)

	Detection ($\beta = 10$; $\tau_2 = 100$ sec; $T_n = 10°K$)					Communication (reception of information) ($\alpha = 100$; $T_n = 10°K$)							
R \ Δf	10^7	10^8	10^9	10^{10}		1	10	10^2	10^3	10^4	10^5	10^6	10^7
10	$2 \cdot 10^{17}$	$2 \cdot 10^{18}$	$2 \cdot 10^{19}$	$2 \cdot 10^{20}$	$2 \cdot 10^{21}$	$2 \cdot 10^{19}$	$2 \cdot 10^{20}$	$2 \cdot 10^{21}$	$2 \cdot 10^{22}$	$2 \cdot 10^{23}$	$2 \cdot 10^{24}$	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$
10^2	$2 \cdot 10^{19}$	$2 \cdot 10^{20}$	$2 \cdot 10^{21}$	$2 \cdot 10^{22}$	$2 \cdot 10^{23}$	$2 \cdot 10^{21}$	$2 \cdot 10^{22}$	$2 \cdot 10^{23}$	$2 \cdot 10^{24}$	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$
10^3	$2 \cdot 10^{21}$	$2 \cdot 10^{22}$	$2 \cdot 10^{23}$	$2 \cdot 10^{24}$	$2 \cdot 10^{25}$	$2 \cdot 10^{23}$	$2 \cdot 10^{24}$	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$
10^4	$2 \cdot 10^{23}$	$2 \cdot 10^{24}$	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$	$2 \cdot 10^{27}$	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$
10^5	$2 \cdot 10^{25}$	$2 \cdot 10^{26}$	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$	$2 \cdot 10^{29}$	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$
10^6	$2 \cdot 10^{27}$	$2 \cdot 10^{28}$	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$	$2 \cdot 10^{31}$	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$	$2 \cdot 10^{35}$	$2 \cdot 10^{36}$
10^7	$2 \cdot 10^{29}$	$2 \cdot 10^{30}$	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$	$2 \cdot 10^{33}$	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$	$2 \cdot 10^{35}$	$2 \cdot 10^{36}$	$2 \cdot 10^{37}$	$2 \cdot 10^{38}$
10^8	$2 \cdot 10^{31}$	$2 \cdot 10^{32}$	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$	$2 \cdot 10^{35}$	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$	$2 \cdot 10^{35}$	$2 \cdot 10^{36}$	$2 \cdot 10^{37}$	$2 \cdot 10^{38}$	$2 \cdot 10^{39}$	$2 \cdot 10^{40}$
10^9	$2 \cdot 10^{33}$	$2 \cdot 10^{34}$	$2 \cdot 10^{35}$	$2 \cdot 10^{36}$	$2 \cdot 10^{37}$	$2 \cdot 10^{35}$	$2 \cdot 10^{36}$	$2 \cdot 10^{37}$	$2 \cdot 10^{38}$	$2 \cdot 10^{39}$	$2 \cdot 10^{40}$	$2 \cdot 10^{41}$	$2 \cdot 10^{42}$

The entire energy consumed is eventually degraded to heat and then radiated into outer space. In principle, it could be convered into radio waves (this does not clash with the thermodynamic laws) and then used for interstellar radio communication. However, the entire power would not be enough for ensuring continuous isotropic transmission aimed at an undirectional receiving antennas within the range of a few tens of light years, i. e., the message would not reach the nearest stars. This does not mean, however, that this convenient method of communication is completely hopeless. Since we assume that our civilization is not unique in the Universe, it inevitably follows that there should be civilizations on a lower technological level than ours, on the same level with us, and of course on higher levels of development. The highly advanced civilizations may have tremendous power resources at their disposal, which are absolutely inaccessible to mankind at the present stage of development. The power potential of a civilization in the last analysis determines the power of its transmitters. On the other hand, this is one of the most important parameters of interstellar communication affecting the range of detection and communication, the quantity of transmitted information, the kind of signals, used, and, indirectly, the methods of detection. Therefore, the question of the probable power potential of a civilization merits a more detailed examination.

The main features of the growth of the principal indices of technological progress of civilizations are analyzed in Chapters I, V, and VI. We will consider here only the growth of the power resources of a civilization.

The annual growth of power consumption in the world is about 3% during the last 100 years. If the same rate of growth will persist in the future, the per-second power consumption on the Earth will reach 10^{17} watt in the next 300 years, thus becoming equal to the influx of solar energy. Further increase of power consumption will be unfeasible, since this will radically

change the radiation balance of the planet.* This is a highly significant factor, whose importance is generally underestimated. It should be emphasized that this restriction of power output has nothing to do with shortage of power resources: it follows from the necessity to maintain the equilibrium in the atmosphere and on the surface of the Earth.

A similar situation is encountered by any civilization on any planet. Since the energy received by a planet in the "life zone" from its primary cannot change between very wide limits, the energy output of any planetary civilization should be limited by figures of the order of 10^{17} watt. When this limit is reached, further development is possible only through active conquest and population of interplanetary space, where high-power installations and industries should be moved. In the light of this conclusion, it seems that the exploration of space which has recently begun is a vital step toward ensuring the future growth and existence of our civilization, and by no means can it be regarded as premature. Active conquest and population of outer space will eventually lead to the creation of an artificial biosphere around the Sun (the Dyson – Tsiolkovskii sphere). A civilization of this kind, inhabiting an artificial biosphere around its primary, should have access to much higher power outputs, reaching $3 \cdot 10^{26}$ watt. Assuming an exponential growth, the time to reach a Dyson-type civilization is not very long. Indeed, if the annual growth of power output is merely 1%, the transition from a civilization with an energy output of 10^{17} watt to a Dyson civilization with energy requirements of the order of $3 \cdot 10^{26}$ watt should take about 2200 years. After another 2500 years, the per-second power consumption, assuming the same growth rate, will reach 10^{37} watt, which is equal to the radiation energy of all the stars in the Galaxy.** This extrapolation of the growth of civilizations can be extended into the more distant future, but we had better stop here. In 1964, Kardashev /10/ proposed a division of all the civilizations into three types in terms of the power requirements. Type I civilizations are those which are close in their technical development to the Earth civilization (power requirements of $10^{12} - 10^{13}$ watt), type II civilizations are those with power requirements of the order of $3 \cdot 10^{26}$ watt, and finally, type III civilizations are those which have harnessed the power resources on a galactic scale, with energy output of 10^{37} watt. We will follow Kardashev's terminology, but extend the concept of type I civilizations to all planetary civilizations with power requirements close to the Earth level and higher, up to 10^{17} watt. The existence of supercivilizations with energy requirements of the order of $10^{26} - 10^{37}$ watt is a mere hypothesis. However, strictly speaking, the very assumption regarding the existence of other extraterrestrial civilizations is also a mere hypothesis. It is therefore advisable not to ignore any of the possibilities.

Let us return to Table 3.2. For a band width of over 1 Hz, detection of signals and reception of information from type I civilizations is ruled out even for the nearest stars. Thus, only type II or type III civilizations can communicate by means of continuous isotropic transmission with undirectional reception. With bands of 1 MHz, the detection of signals from

* Actually, the power output level will have to be frozen at a much earlier stage, when it reaches a few per cent of the solar energy flux received each second (i.e., in about 100–200 years).

** For a higher annual growth rate, these limit values will be attained much sooner (see, e.g., Chapter I, p. 26).

type II civilizations is possible over distances of 1000 light years, but reception of information is possible only over distances corresponding to the nearest stars. For band widths $\Delta f \leqslant 100$ Hz, signals from type II civilizations can be detected anywhere in the Galaxy, whereas reception of information is possible over distances greater than 1000 light years. The signals of type III civilizations can be detected virtually anywhere in the observable Universe. For band widths of 10 MHz, information can be transmitted only to the nearest galaxies, whereas for sufficiently narrow bands, $\Delta f \leqslant 100$ Hz, information can be transmitted within the limits of the Metagalaxy.

Hence it follows that if at least one type II civilization exists anywhere in our Galaxy or at least one type III civilization exists anywhere in the Universe, and these civilizations using their tremendous power potential transmit continuously in all directions powerful monochromatic signals of band width $\Delta f \leqslant 100$ Hz, we should be able to detect these signals even without knowing where the source is located.

This method of communication is the least advantageous in terms of power. Other methods of transmission and reception require much lower power levels. Consider a high-quality receiver with noise temperature of 10°K and band width of 100 MHz which functions at 3 cm wavelength. Suppose that this receiver is placed outside the atmosphere, where the total noise temperature is determined by the receiver noises, being equal to $T_n = 10$°K. Let us now determine what power is needed for the detection of signals or reception of information over distances of 1000 light years for various reception and transmission techniques. We have chosen the distance of 1000 light years because, according to some modern estimates /2, 3, 15/, this is the average distance to the nearest EC. Table 3.3 lists the power values (in watt) necessary for signal detection and communication over distances of 1000 light years assuming $\Delta f_1 < \Delta f_2$. An area of 10^4 m^2 was assumed for the receiving antenna, which corresponds to the area of the largest modern radio telescopes in the centimeter and decimeter range. For directional transmission, the table gives the product of power times the antenna gain in dB · W and also the power corresponding to the gain $g_1 = 10^9$. The dashes in the last two columns indicate that, formally, the condition of pulse signal detection ($\tau_2 < \Delta t_1$) coincides with the condition of information reception (see §2). Note that for the problem of communication with EC, the main factor is not the relative pulse duration $\frac{\Delta t_1}{t_1}$, but the length of time t_1 between the successive pulses, which determines the rate of information transmission by pulse signals. Since in our case $\Delta t_1 = \Delta f_1^{-1} > 10^{-8}$, the relative pulse durations listed in the table correspond to the following information transmission rates:*

$$\frac{\Delta t_1}{t_1} = 10^{-1},\ 10^{-2},\ 10^{-3},\ 10^{-4},\ 10^{-6},\ 10^{-8},\ 10^{-10}$$

$$t_1 > 10^{-7},\ 10^{-6},\ 10^{-5},\ 10^{-4},\ 10^{-2},\ 1,\ 10^2 \text{ sec}$$

$$q = \frac{1}{t_1} < 10^7,\ 10^6,\ 10^5,\ 10^4,\ 10^2,\ 1,\ 10^{-2} \text{ bit/sec.}$$

* On the assumption that binary pulses are used. If a pulsed code with some base $a \neq 2$ is used, the transmission rate figures should be multiplied by $\log_2 a$.

TABLE 3.3. Transmitter power needed for the detection of narrow band signals and communication over distances of 1000 light years ($\lambda = 3$ cm, $\Delta f_1 < \Delta f_2 = 100$ MHz, $T_n = 10°K$)

Transmission	Communication; $\alpha = 100$				Detection; $\beta = 10$, $\tau_2 = 100$ sec			
	nondirectional reception $T_n = 10°K$		directional reception $S_2 = 10^4$ m^2; $T_n = 10°K$		nondirectional reception $T_n = 10°K$		directional reception $S_2 = 10^4$ m^2; $T_n = 10°K$	
Isotropic continuous	$2 \cdot 10^{31}$		$1.5 \cdot 10^{23}$		$2 \cdot 10^{25}$		$1.5 \cdot 10^{17}$	
Isotropic pulse $\frac{\Delta t_1}{t_1} =$								
10^{-1}	$2 \cdot 10^{30}$		$1.5 \cdot 10^{22}$					
10^{-2}	$2 \cdot 10^{29}$		$1.5 \cdot 10^{21}$					
10^{-4}	$2 \cdot 10^{27}$		$1.5 \cdot 10^{19}$					
10^{-6}	$2 \cdot 10^{25}$		$1.5 \cdot 10^{17}$		—		—	
10^{-8}	$2 \cdot 10^{23}$		$1.5 \cdot 10^{15}$					
10^{-10}	$2 \cdot 10^{21}$		$1.5 \cdot 10^{13}$					
Directional continuous	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$
	313	$2 \cdot 10^{22}$	232	$1.5 \cdot 10^{14}$	253	$2 \cdot 10^{16}$	172	$1.5 \cdot 10^{8}$
Directional pulse $\frac{\Delta t_1}{t_1} =$								
10^{-1}	303	$2 \cdot 10^{21}$	222	$1.5 \cdot 10^{13}$				
10^{-2}	293	$2 \cdot 10^{20}$	212	$1.5 \cdot 10^{12}$				
10^{-4}	273	$2 \cdot 10^{18}$	192	$1.5 \cdot 10^{10}$				
10^{-6}	253	$2 \cdot 10^{16}$	172	$1.5 \cdot 10^{8}$	—		—	
10^{-8}	253	$2 \cdot 10^{14}$	152	$1.5 \cdot 10^{6}$				
10^{-10}	213	$2 \cdot 10^{12}$	132	$1.5 \cdot 10^{4}$				

We see from Table 3.3 that type I civilizations will remain undetected at a distance of 1000 light years in the case of continuous isotropic transmission and nondirectional reception. Signals from type II civilizations can be detected under these conditions, but there is not enough power for the reception of information. To ensure information reception, we should switch over either to pulse or to directional transmission or, alternatively, to directional reception. In the case of continuous isotropic reception and directional reception, we can detect signals from type I civilizations and receive information from type II civilizations over these distances. Isotropic pulse transmission with relative pulse duration $\left(\frac{\Delta t_1}{t_1}\right) < 10^{-5}$ makes it possible to establish communication with type II civilizations using a nondirectional receiving antenna. The transition to a directional receiving antenna with an effective area of 10^4 m^2 makes it possible to establish communication with type II civilizations for almost any relative pulse duration. For $\left(\frac{\Delta t_1}{t_1}\right) \leqslant 10^{-6}$, information can be received from type I civilizations (at transmission rates lower than 100 bit/sec); in particular, for $\left(\frac{\Delta t_1}{t_1}\right) = 10^{-10}$, when the rate of information transmission is higher than 1 bit/min, power of the order of 10^{13} watt is required, and this figure is

comparable with the present-day power output of mankind. In the case of continuous directional transmission and nondirectional reception, we can detect signals from type I civilizations and receive information from type II civilizations. In case of directional reception and transmission, 10^{11} kW are required for the reception of information and 150 MW for the detection of signals. Directional pulse transmission makes it possible to establish communication with a type II civilization using a nondirectional receiving antenna for pulses of any relative duration. For relative pulse duration of less than 10^{-5}, communication with type I civilizations is possible if a nondirectional receiving antenna is used; with a directional antenna having an effective area of 10^4 m^2, communication with type I civilizations is possible for any pulse duration (continuous transmission included). Finally, in case of continuous pulse transmission with relative pulse duration of less than 10^{-8} and reception with a directional antenna of effective area of 10^4 m^2, a mere 1.5 MW is needed

Table 3.4 lists the power values required for the detection of signals with a communication system with the same parameters assuming $\Delta f_1 > \Delta f_2$. Note that in this case we can only discuss signal detection, since for $\Delta f_1 > \Delta f_2$ communication inevitably involves signal distortion and loss of information. This should be kept in mind in reference to the left half of the table, which lists the power values required for this incomplete communication. In distinction from the previous case ($\Delta f_1 < \Delta f_2$), the power in pulse transmission now depends on the time spacing between the pulses, and not on the relative pulse duration. When comparing the data of Table 3.4 with the previous data of Table 3.3, we should remember that since now $\Delta t_1 = \Delta f_1^{-1} < 10^{-8}$ the values of t_1 listed in Table 3.4 correspond to the following relative pulse durations:

$t_1 =$	1^s	1^m	1^h	24^h	1 month	1 year
$\frac{\Delta t_1}{t_1} <$	10^{-8}	$2 \cdot 10^{-10}$	$3 \cdot 10^{-12}$	10^{-13}	$4 \cdot 10^{-15}$	$3 \cdot 10^{-16}$

Table 3.5 lists the minimum power values required for communication over distances of 1000 light years assuming equal receiver and transmitter band widths. Examining this table, we readily see that in case of directional reception and transmission, detection of signals and communication can be established over distances of 1000 light years with very small power, especially if pulse transmission is used. From the point of view of power requirements, this is the best method of communication. However, the detection of these signals is very unlikely, unless the direction of transmission and reception are known in advance. Isotropic transmission should be used, as we have noted before, to enable new subscribers to tune in. If the transmission is continuous in time, it is better to use directional receiving antennas subsequently scanning different parts of the sky. This procedure requires power of the order of $10^{13} - 10^{17}$ watt, which is available to type I civilizations. In case of pulse signals, especially when the pulses follow one another at large intervals, it is better to use a nondirectional antenna, since in this way a continuous sky survey can be conducted and the probability of detection markedly increases. This approach, however, requires power of the order of $10^{15} - 10^{23}$ watt. The lower of these figures corresponds to very slow transmission (1 bit per year), so that we should actually speak of signal detection, and not information reception. (There is a possibility, however, that transmission

rates of the order of 1 bit per hour or even 1 bit per year are acceptable for call signals.) An even better way to search for pulse signals is with the aid of a system of directional antennas, which jointly cover the entire sky. If each antenna has an area $S_2 = 10^4\ m^2$, power of $10^7 - 10^{15}$ watt will suffice for the detection of pulse signals.

TABLE 3.4. Transmitter power needed for the detection of wide-band signals in communication (with partial loss of information) over distances of 1000 light years ($\lambda = 3$ cm, $\Delta f_1 > \Delta f_2 = 10^8$ Hz, $T_n = 10°K$)

No.	Transmission	Communication (with loss of information) $\alpha = 100$				Detection; $\beta = 10$; $\tau_2 = 100$ sec			
		nondirectional reception, $T_n = 10°K$		directional reception, $S_2 = 10^4 m^2$; $T_n = 10°K$		nondirectional reception, $T_n = 10°K$		directional reception, $S_2 = 10^4 m^2$; $T_n = 10°K$	
1	Isotropic continuous								
	$\Delta f_1 = 10^9$		$2 \cdot 10^{32}$		$1.5 \cdot 10^{24}$		$2 \cdot 10^{26}$		$1.5 \cdot 10^{18}$
	$\Delta f_1 = 10^{10}$		$2 \cdot 10^{33}$		$1.5 \cdot 10^{25}$		$2 \cdot 10^{27}$		$1.5 \cdot 10^{19}$
2	Isotropic pulse $t_1 =$								
	1 sec		$2 \cdot 10^{23}$		$1.5 \cdot 10^{15}$				
	1 min		$3 \cdot 10^{21}$		$2.5 \cdot 10^{13}$				
	1 hr		$6 \cdot 10^{19}$		$4 \cdot 10^{11}$		–		–
	24 hr		$2 \cdot 10^{18}$		$2 \cdot 10^{10}$				
	1 month		$8 \cdot 10^{16}$		$6 \cdot 10^{8}$				
	1 year		$7 \cdot 10^{15}$		$5 \cdot 10^{7}$				
3	Directional continuous	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$
	$\Delta f_1 = 10^9$	323	$2 \cdot 10^{23}$	242	$15 \cdot 10^{15}$	263	$2 \cdot 10^{17}$	182	$1.5 \cdot 10^{9}$
	$\Delta f_1 = 10^{10}$	333	$2 \cdot 10^{24}$	252	$15 \cdot 10^{16}$	273	$2 \cdot 10^{18}$	192	$1.5 \cdot 10^{10}$
4	Directional pulse $t_1 =$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$	$P_1 g_1$ dB · W	P_1 for $g_1 = 10^9$				
	1 sec	233	$2 \cdot 10^{14}$	152	$1.5 \cdot 10^{6}$				
	1 min	215	$3 \cdot 10^{12}$	134	$2.5 \cdot 10^{4}$				
	1 hr	197	$6 \cdot 10^{10}$	116	$4 \cdot 10^{2}$		–		–
	24 hr	184	$2 \cdot 10^{9}$	103	$4 \cdot 10^{-1}$				
	1 month	169	$8 \cdot 10^{7}$	88	$6 \cdot 10^{-2}$				
	1 year	158	$7 \cdot 10^{6}$	77	$6 \cdot 10^{-2}$				

Once two civilizations have discovered each other, they may establish bilateral directional communication.* In this case, information can be transmitted at a rate of $1 - 10^8$ bits/sec over distances of 1000 light years with power of the order of $10^6 - 10^{14}$ watt, i.e., substantially less than the power required for signal detection. This, apparently paradoxical, conclusion is quite understandable: the high power needed for signal detection is the price we have to pay for not knowing the subscriber's address.

* The concept of bilaterally directional radio communication does not refer to a dialogue between civilizations (no such dialogue is possible in interstellar communication, since the answer will take thousands of years to cross the tremendous distance), but rather the two unilateral "monologues" transmitted through a channel with a directional transmitting antenna and a directional receiving antenna.

TABLE 3.5. Minimum power needed for signal detection and reception of information over distances of 1000 light years ($\Delta f_1 = \Delta f_2 = \Delta f$; $\lambda = 3$ cm)

No.	Transmission	Reception of information $\alpha = 100$		Detection; $\beta = 10$; $\tau_1 = 100$ sec	
		nondirectional reception, $T_n = 10°K$	directional reception, $S_1 = 10^4 m^2$; $T_n = 10°K$	nondirectional reception, $T_n = 10°K$	directional reception, $S_1 = 10^4 m^2$; $T_n = 10°K$
1	Isotropic continuous $\Delta f =$				
	1	$2 \cdot 10^{23}$	$1.5 \cdot 10^{15}$	$2 \cdot 10^{21}$	$1.5 \cdot 10^{13}$
	10^2	$2 \cdot 10^{25}$	$1.5 \cdot 10^{17}$	$2 \cdot 10^{22}$	$1.5 \cdot 10^{14}$
	10^4	$2 \cdot 10^{27}$	$1.5 \cdot 10^{19}$	$2 \cdot 10^{23}$	$1.5 \cdot 10^{15}$
	10^6	$2 \cdot 10^{29}$	$1.5 \cdot 10^{21}$	$2 \cdot 10^{24}$	$1.5 \cdot 10^{16}$
	10^8	$2 \cdot 10^{31}$	$1.5 \cdot 10^{23}$	$2 \cdot 10^{25}$	$1.5 \cdot 10^{17}$
2	Isotropic pulse $t_1 =$				
	1 sec	$2 \cdot 10^{23}$	$1.5 \cdot 10^{15}$		
	1 min	$3 \cdot 10^{21}$	$2.5 \cdot 10^{13}$		
	1 hr	$6 \cdot 10^{19}$	$4 \cdot 10^{11}$	–	–
	24 hr	$2 \cdot 10^{18}$	$2 \cdot 10^{10}$		
	1 month	$8 \cdot 10^{16}$	$6 \cdot 10^{8}$		
	1 year	$7 \cdot 10^{15}$	$5 \cdot 10^{7}$		
3	Directional continuous $\Delta f =$	$P_1 g_1$ dB·W / P_1 for $g_1 = 10^9$	$P_1 g_1$ dB·W / P_1 for $g_1 = 10^9$	$P_1 g_1$ dB·W / P_1 for $g_1 = 10^9$	$P_1 g_1$ dB·W / P_1 for $g_1 = 10^9$
	1	233 / $2 \cdot 10^{14}$	152 / $1.5 \cdot 10^{6}$	213 / $2 \cdot 10^{12}$	132 / $1.5 \cdot 10^{4}$
	10^2	253 / $2 \cdot 10^{16}$	172 / $1.5 \cdot 10^{8}$	223 / $2 \cdot 10^{13}$	142 / $1.5 \cdot 10^{5}$
	10^4	273 / $2 \cdot 10^{18}$	192 / $1.5 \cdot 10^{10}$	233 / $2 \cdot 10^{14}$	152 / $1.5 \cdot 10^{6}$
	10^6	293 / $2 \cdot 10^{20}$	212 / $1.5 \cdot 10^{12}$	243 / $2 \cdot 10^{15}$	162 / $1.5 \cdot 10^{7}$
	10^8	313 / $2 \cdot 10^{22}$	232 / $1.5 \cdot 10^{14}$	253 / $2 \cdot 10^{16}$	172 / $1.5 \cdot 10^{8}$
4	Directional pulse $t_1 =$				
	1 sec	333 / $2 \cdot 10^{14}$	152 / $1.5 \cdot 10^{6}$		
	1 min	215 / $3 \cdot 10^{12}$	134 / $2.5 \cdot 10^{4}$		
	1 hr	197 / $6 \cdot 10^{10}$	116 / $4 \cdot 10^{2}$	–	–
	24 hr	184 / $2 \cdot 10^{9}$	103 / $2 \cdot 10$		
	1 month	169 / $8 \cdot 10^{7}$	88 / $6 \cdot 10^{-1}$		
	1 year	158 / $7 \cdot 10^{6}$	77 / $5 \cdot 10^{-2}$		

Let us now try to establish the cost of interstellar communication. Consider two civilizations at a distance of 1000 light years from each other which have established bilaterally directional radio communication using a high-frequency channel of 1 Hz band width. From Table 3.5 we find the transmitter power needed for this communication; it is 1000 kW, which is easily accessible to civilizations on our level. The signal-to-noise ratio at the reception point will then be 100, and this is again quite adequate for establishing reliable radio communication with the aid of binary PCM. The transmission rate of PCM information through a 1 Hz wide channel is 1 bit/sec. Let us find the cost of a 100 word transmission through this channel. The message is composed using a 30-letter alphabet and each word contains on the average five letters. Our message then contains $2.5 \cdot 10^3$ bits of information (see § 1) and if transmitted at a rate of 1 bit/sec, it will take $2.5 \cdot 10^3$ sec. Assuming a transmitter power of 1000 kW, the message will consume about 700 kW-hr of energy, costing about 28 rubles. We see that the communication with extraterrestrial civilizations is not very expensive. The main problem is to discover your counterpart.

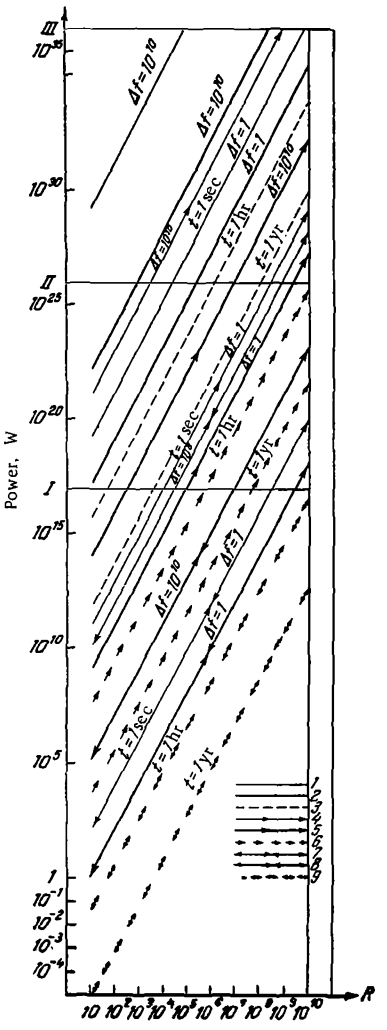


FIGURE 45. *Minimum transmitter power* ($\Delta f_2 = \Delta f_1 = \Delta f$) *vs. communication range at centimeter wavelengths for various methods of reception and transmission.*

1 — range of communication for isotropic transmission and nondirectional reception; 2 — range of detection for isotropic transmission and nondirectional reception; 3 — isotropic transmission of pulse signals, nondirectional reception; 4 — range of communication for isotropic transmission and directional reception, effective area of receiving antenna $S_1 = 10^4$ m^2; 5 — range of detection for isotropic transmission and directional reception; 6 — isotropic transmission of pulse signals, directional reception; 7 — range of communication for directional transmission and directional reception, $g_1 = 10^9$, $S_1 = 10^4$ m^2; 8 — range of detection for directional transmission and reception; 9 — directional transmission of pulse signals, directional reception; t is the time spacing between the pulses. The calculations were made for $\lambda = 3$ cm, $\alpha = 100$, $\beta = 10$, $T_n = 10^2$°K.

Figure 45 plots the minimum transmission power vs. the range of communication for various methods of reception and transmission. In the top left corner of the figure lies the region of isotropic transmission of wide-band signals with nondirectional reception. In the bottom right corner we have the region corresponding to directional transmission of narrow-band or pulse signals and directional reception. All the intermediate cases fall in between.

It is left to the reader to choose on the ordinate axis the power values corresponding to his skepticism or imagination, and to determine the optimum methods of transmission and reception for any distance in the Universe.

Let us try to estimate the possibility of detection of EC which do not send special signals (by "listening in" to their internal radio transmissions). The power involved in these transmissions will be of the order of $10^5 - 10^6$ watt. At this power level, the isotropic wide-band ultra-short-wave transmissions cannot be detected even from the nearest stars. Interception of highly directional radio transmissions with a directivity coefficient of the order of 10^9, which extraterrestrial civilizations may use for some special purposes (e.g., interplanetary communication), is possible over distances of a few hundreds and even thousands of light years. However, the probability that such a tight message will be accidentally intercepted by the receiving antenna is very low. The planetary rotation increases this probability, but it nevertheless remains low, if we remember that at a distance of 100 — 1000 light years there are less than ten transmitting EC. Thus detection of extraterrestrial civilizations with the aid of their routine radio communications is virtually impossible. To become detectable, they must transmit special signals in the form of powerful isotropic radiation or in the form of highly directional radiation with a scanning antenna.*

Radio communication between galaxies

Let us consider some specific features of intergalactic radio communication. There is probably at least one civilization per galaxy capable of transmitting and receiving information, so that it is worth trying to probe the individual galaxies with directional receiving and transmitting antennas. The antenna directivity should be chosen so that the beam would cover the entire galaxy, i.e., so that $\varphi = \theta$, where φ is the angular size of the galaxy, and θ is the beam width. Using relation (3.55) between beam width and directivity coefficient and the dependence of the angular size of a galaxy on its distance, we may write this condition in the form

$$\frac{L^2}{R^2} = \frac{10}{g}, \tag{3.69}$$

where L is the mean linear size of a galaxy. Suppose a civilization situated in a certain galaxy sends an isotropic transmission ($g_1 = 1$) and

* From the point of view of reception and power requirements, this system is equivalent to isotropic pulse transmission.

another civilization in a nearby galaxy uses a directional receiving antenna of gain g_2 satisfying (3.69). Inserting the corresponding values of g_1 and g_2 in (3.33c), we obtain an expression of the power needed for this communication:

$$P_1 = 16\alpha k T_n \Delta f \left(\frac{L}{\lambda}\right)^2. \tag{3.70}$$

This is a paradoxical result, because the required power turns out to be independent of distance! The same conclusion is obtained if some EC sends a transmission directed at another galaxy with an antenna which completely covers the recipient $\left(g_1 = 10\,\frac{R^2}{L^2}\right)$, whereas the civilization in the receiving galaxy uses a nondirectional antenna ($g_2 = 1$). Let us calculate the numerical value of the power needed for this kind of intergalactic communication. Setting in (3.70) $\alpha = 100$, $T_n = 10°K$, $\lambda = 3$ cm, $\Delta f = 1$ Hz, we find $P_1 = 2 \cdot 10^{26}$ watt.

In case of bilaterally directional communication between the galaxies, the receiving and the transmitting antenna gains satisfy (3.69). The power needed for this communication is therefore

$$P_1 = \frac{16\alpha k T_n \Delta f L^4}{10\lambda^2 R^2}. \tag{3.71}$$

This result is even more puzzling: the greater the distance between the communicating galaxies, the lower is the power needed for establishing the communication!

Troitskii /6/ was the first to call attention to these peculiar features of intergalactic communication. They seem to stem from condition (3.69), i.e., they are basically associated with the fact that the antenna directivity is a function of the angular dimensions of the galaxy (the area of each antenna has to be increased in proportion to the intergalactic distance). The antenna parameters and the power required for intergalactic communication emerge from Table 3.6, which lists the directivity coefficients, the receiving antenna areas, and the power for communication with two neighbor galaxies (the Large Magellanic Cloud and the Andromeda Nebula) and with some other typical galaxy with dimensions of the order of 10^5 light years. Here, as before, $\alpha = 100$, $T_n = 10°K$, $\Delta f_1 = \Delta f_2 = \Delta f = 1$ Hz.

TABLE 3.6. The directivity coefficient of antennas and the required power for intergalactic communication

Galaxies	Large Magellanic Cloud	Andromeda	A typical galaxy with $L = 10^5$ light years			
Distance in light years .	$2 \cdot 10^5$	$2 \cdot 10^6$	10^7	10^8	10^9	10^{10}
Angular dimensions . . .	9°	3°.5	34'	206"	21"	2"
Directivity coefficient of receiving antenna . . .	360	$3 \cdot 10^3$	10^5	10^7	10^9	10^{11}
Area (m^2) at $\lambda = 3.5$ cm .	0.04	0.3	10	10^3	10^5	10^7
Transmitter power, watt.	$6 \cdot 10^{23}$	10^{24}	$2 \cdot 10^{22}$	$2 \cdot 10^{20}$	$2 \cdot 10^{18}$	$2 \cdot 10^{16}$

Monochromatic signals. Frequency scanning

We have so far considered the power aspects of communication. Now we can discuss in more detail the various aspects relating to signal band width. Here we should distinguish between two cases: wide-band signals with a virtually continuous spectrum ($\Delta\nu \sim \nu$) and narrow-band monochromatic signals with a band width substantially narrower than the frequency ($\Delta\nu \ll \nu$).

Frequency scanning constitutes one of the basic stages in any search for narrow-band signals. As we have noted before, the band width in low-speed transmission (e.g., in case of call signals) is determined entirely by the stability of the transmitted signal. It may be as low as fractions of Hz. On the other hand, the width of the optimum frequency range where EC signals can be expected is of the order of $10^{10} - 10^{11}$ Hz. Our problem is thus how to detect a narrow line of relative width of $10^{-11} - 10^{-10}$ in this frequency range. Since the direction in which these signals should be sought is not known either, the problem, to borrow Purcell's expression, is not unlike that of trying to meet a certain person in New York City without having previously agreed on a meeting place. Nevertheless, this complex problem is basically and technically solvable.

We will first consider the question of frequency scanning, and then proceed to direction scanning. There are two methods of frequency scanning currently known: a single-channel scanning receiver with automatic frequency tuning or a multichannel receiver with narrow-band filters, each tuned to a certain frequency and all the filters jointly covering the entire frequency range. Which of the two techniques is to be preferred? To answer this question, we have to use some logical evaluation criterion. Siforov /8/ proposed the following criterion. In order to detect EC signals, we have to ensure reception of a sufficient quantity of information which will indicate with high reliability the artificial origin of the radio source. It is therefore best to use those signal detection methods which provide the essential minimum of information in minimum time. Let us now evaluate the two frequency scanning techniques from this point of view.

Consider a single-channel scanning receiver with continuous frequency tuning, the block diagram of which is shown in Figure 46.

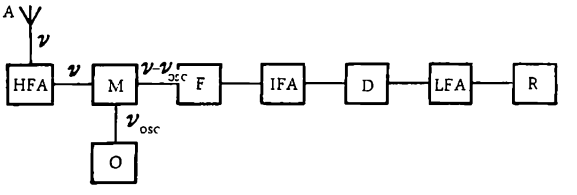


FIGURE 46. Block diagram of a single-channel frequency-scanning receiver:

A — antenna, HFA — high frequency amplifier, O — oscillator, M — mixer, F — filter, IFA — intermediate-frequency amplifier, D — detector, LFA — low-frequency amplifier, R — recorder.

The oscillator frequency ν_{osc} follows the frequency of the tunable HF amplifier. The mixer transforms the instantaneous high frequency ν into a constant intermediate frequency $\nu_{int}=\nu-\nu_{osc}$ After the mixer, the signal is passed through an intermediate-frequency filter, amplified, detected, and transmitted to a recorder through a low-frequency amplifier. The receiver band width Δf_2 is limited by the intermediate-frequency filter, $\Delta f_2=\Delta f_{int}$. Let Δt be the time for the current frequency to scan a frequency band Δf_2. As this band width is being scanned, the filter input receives a certain signal $x(t)$ of duration Δt whose band width is $\Delta f=\Delta t^{-1}$. To ensure undistorted transmission of this signal through the filter, we should clearly have $\Delta f\leqslant\Delta f_2$ or $\Delta t\geqslant\Delta f_2^{-1}$. Hence it follows, that the rate of frequency variation in this tunable system depends on the receiver band width Δf_2:

$$\frac{d\nu}{dt}=\frac{\Delta f_2}{\Delta t}\leqslant(\Delta f_2)^2. \tag{3.72}$$

This rate of tuning is limited from above, and the time of scanning of a given frequency band therefore cannot be made as small as desired. The maximum rate of frequency variation is

$$\left(\frac{d\nu}{dt}\right)_{max}=(\Delta f_2)^2, \tag{3.73}$$

and the corresponding minimum scanning time for a frequency band Δf_2 is

$$\Delta t_{min}=\Delta f_2^{-1}. \tag{3.74}$$

Let Δf_0 be the scanned frequency range. The total scanning time required for the current tunable frequency to run through the entire relevant range is thus

$$\tau_{int}=N\,\Delta t_{min}=\frac{\Delta f_0}{(\Delta f_2)^2}, \tag{3.75}$$

where N is the number of elementary frequency bands, i.e., receiver bands Δf_2, accommodated in the scanned frequency range. We thus see that the time of frequency scanning is inversely proportional to the square of the band width of a single-channel scanning receiver. This is quite understandable, since a decrease of the receiver band width, on the one hand, increases the number of elementary bands into which the scanned frequency range is divided and, on the other, increases the scanning time in each elementary band (since the rate of frequency variation decreases in proportion to the square of the band width).

Let us now express the scanning time as a function of the distance to the source. Since $\Delta f_2\propto R^{-2}$, equation (3.75) may be written in the form

$$\tau_{int}\propto\Delta f_0 R^4. \tag{3.76}$$

Thus, for the particular reception technique using a single-channel scanning receiver, the time of search for monochromatic EC signals is seen to be directly proportional to the total frequency band Δf_0 in which the search is conducted and to the fourth power of the distance to the sending civilization.

If we use a multichannel receiver made up of N channels of band width Δf_2 each, all the channels fully covering the required frequency range ($N\Delta f_2 = \Delta f_0$), the total frequency scanning time will be equal to the time to scan a single channel. It follows from (3.74) that in this case the time is inversely proportional to the band width of each channel or, using the relationship between range and band width, it is proportional to the square of the distance between civilizations.

Consider a search for signals at 3 cm wavelength when the distance to the sending civilization is 1000 light years, the transmitter power is 150 MW, the transmitting antenna gain is 90 dB, the effective area of the receiving antenna is 10^4 m^2, and the noise temperature is 10°K. Suppose that information can be received for signal-to-noise ratios of 100. We see from Table 3.5 that the receiver band width in this case should be 100 Hz. The minimum scanning time for this band width is of the order of 0.01 sec. If a single-channel scanning receiver is used, several years will be needed to scan the entire frequency range around 3 cm ($\Delta f_0 = 10^{10}$ Hz). A multichannel receiver comprising 10^8 channels each 100 Hz high will scan the entire frequency range in a time of the order of 0.01 sec. If the distance to the sending civilization is increased 10-fold, the scanning time with a multichannel receiver will increase 100-fold reaching 1 sec, whereas for a single-channel scanning receiver the time will increase by a factor of 10^4, reaching 30,000 years.

We thus see that the scanning time of a continuously tunable single-channel receiver is much longer than the scanning time of a multichannel receiver. Moreover, as the distance between the civilizations increases, the scanning time with a single-channel receiver grows much faster ($\tau_{int} \propto R^4$) than the scanning time of a multichannel system ($\tau_{int} \propto R^2$).

All this renders single-channel scanning receivers practically useless for the detection of monochromatic signals from extraterrestrial civilizations. This problem can be tackled more successfully using multichannel receivers with a great number of narrow-band filters. It should be kept in mind that the reduction of scanning time is accomplished as a result of a much greater complexity of the receiving equipment, the instrumental complexity (the number of channels in the receiver) increasing in proportion to R^2. Nevertheless, it seems that this complexity is not without its advantages /8/, since band filters (and other components used in multichannel systems) are cheap and readily accessible.

A multichannel system specifically designed for the detection of monochromatic EC signals was proposed by Kotel'nikov /16/. A block diagram of the receiver is shown in Figure 47. Here A is the antenna, AM is the amplifier which also transforms the frequency of the incoming signals, if necessary, F are filters of band width Δf, covering jointly the entire frequency range, D are the detectors, I are the integrators integrating the energy which passes through the filter in a time τ, T are the threshold devices which produce an output signal only if the energy transmitted through the filter in a time τ exceeds a certain threshold value.

This receiver clearly cannot be used in reception of information transmitted by one of the amplitude modulation techniques. However, information may be sent by varying the frequency of the signal from one transmission to the next. In this case, a signal will appear in one of the receiver channels, and every successive transmission will be picked up by a different channel.

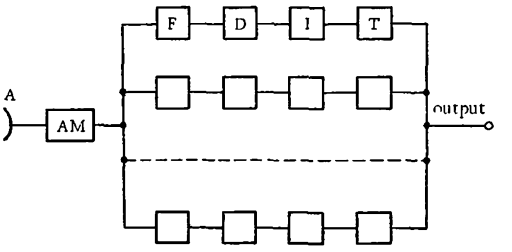


FIGURE 47. Kotel'nikov's multichannel frequency-scanning receiver:

A — antenna, AM — amplifier, F — narrow-band filters, D — detectors, I — integrators, T — threshold elements.

The appearance of a signal in a new channel may be regarded as a certain message. Since one out of N possible channels is selected, each of these messages contains $\log_2 N$ information units (see §1). The rate of information transmission of this system is thus

$$q = \frac{\log_2 N}{\tau}. \tag{3.77}$$

Let us now find the range of communication. For bilaterally directional radio communication, the range of signal reception with Kotel'nikov's receiver is

$$R = \left(\frac{P_1 g_1 S_2 \tau}{4\pi k T_n \Psi} \right)^{1/2}. \tag{3.78}$$

228

The factor Ψ in the denominator depends on the number of channels N, the band width Δf of each channel, and the particular threshold used, which in its turn is determined by the probability of a false response $p_{f.r.}$ of the system (as a result of random noise in the receiver) and the probability of missing a signal p_{miss}. The range of communication increases with the decrease in Ψ. The minimum value of Ψ is attained for $\Delta f = \frac{1}{\tau}$; it is equal to

$$\Psi_{min} = \left(\sqrt{\ln \frac{N}{p_{f.r.}} - 2} + \sqrt{\ln \frac{1}{p_{miss}} - 2} \right)^2. \tag{3.79}$$

This Ψ for a given τ gives the maximum range of communication:

$$R_{max} = \left(\frac{P_1 g_1 S_2}{\Psi_{min} 4\pi k T_n \Delta f} \right)^{1/2}. \tag{3.80}$$

This expression is analogous to (3.33). The only difference is that the α in the numerator has been replaced with Ψ_{min}; its meaning is the same as that of α, since it is determined by the system threshold. It follows from the last two relations that the maximum range of communication falls with an increase in the number of channels and an increase in the

band width of each channel. If a certain frequency band Δf_0 is to be scanned, we have $\Delta f = \frac{\Delta f_0}{N}$. As N increases, the frequency band of each channel decreases faster than Ψ_{min} increases; as a result, the maximum range of communication increases with the increase in the number of channels. For $\Delta f \gg \frac{1}{\tau}$, we have

$$\Psi = (2\Psi_{min}\tau\,\Delta f)^{1/2}, \tag{3.81}$$

and for a given frequency interval Δf_0, the range of communication is

$$R = \left(\frac{P_1 g_1 S}{4\pi k T_n}\sqrt{\frac{\tau N}{2\Psi_{min}\,\Delta f_0}}\right)^{1/2}. \tag{3.82}$$

Let us consider one example. An EC transmitter of 100 MW power sends monochromatic signals in the 3 cm wavelength range in the form of pulses of 100 sec duration, varying the pulse frequency from one transmission to the next. The transmitting antenna has a gain of 10^9, and the reception is carried out with an antenna of 10^4 m^2 area and Kotel'nikov's receiver with $N = 10^9$ channels, noise temperature $T_n = 10°$K, integration time $\tau = 100$ sec, channel width 1 Hz; the false response probability and the signal omission probability are 10^{-5}. Inserting these values of N, τ, Δf, $p_{f.r.}$, and p_{miss} in (3.79) and (3.81), we find $\Psi = 120$. Equation (3.78) then gives the range $R = 7 \cdot 10^4$ light years. The transmission rate of this communication system is 0.01 $\log_2 10^9$ = 0.3 bit/sec. As τ decreases, the range of communication slowly diminishes, whereas the transmission rate increases fairly rapidly. If we take in our example $\tau = 1$ sec, we find $\Psi = \Psi_{min} = 73$, $R = 9 \cdot 10^3$ light years, and the transmission rate will increase to 30 bit/sec.

It is interesting to compare these numbers with the corresponding data for a single-channel receiver operating at a fixed frequency. Using Figure 45 we find that for $P_1 = 10^8$ watt, $g_1 = 10^9$, $S_2 = 10^4$ m^2, $T_n = 10°$K, $\Delta f = 1$ Hz, and $\alpha = 100$ the range of bilaterally directed communication is $8 \cdot 10^3$ light years and the transmission rate is 1 bit/sec, i.e., 1/30 of the transmission rate attainable with Kotel'nikov's system for the same range of communication. The range of detection for these system parameters and $\beta = 10$, $\tau_2 = 100$ sec is $8 \cdot 10^4$ light years, i.e., of the same order of magnitude as the range of communication attainable with a multichannel receiver with transmission rate of 0.3 bit/sec and equal other parameters.

We thus conclude that Kotel'nikov's multichannel receiver is an optimal system for searching for monochromatic signals when the frequency band to be scanned is not too wide. However, the number of channels required to detect a line narrower than 1 Hz in a frequency band of $10^{10} - 10^{11}$ Hz is uncomfortably large. To avoid this difficulty, Troitskii proposed an original combination method. A special spectrum analyzer is applied to cover simultaneously a sufficiently wide part of the spectrum with band width Δf_0 of the order of 1 MHz. In this way, the presence or absence of a monochromatic sine signal in some frequency range can be immediately established. The exact frequency of the signal is not determined, and only some wide

frequency band Δf_0 containing the signal is identified. Once the relevant frequency interval has been identified, a multichannel receiver is applied to exactly determine the signal frequency. For a channel width of 1 Hz, 10^6 channels are required to cover a band of $\Delta f_0 = 1$ MHz in which a signal has been detected. This is not an excessively large number of channels. Moreover, the construction of a multichannel receiver covering the particular frequency interval will be justified by the detection of a signal in that interval.

This method should be first applied to frequencies near the 21 cm hydrogen line, near its harmonics, near the 18 cm hydroxyl OH line, and also possibly near the 1.25 cm ammonia line and the 0.4 cm formaldehyde line used in molecular masers.

Direction scanning

Let us now consider the direction scanning in the general search for signals. Suppose that the distance to the nearest civilization sending meaningful signals into space does not exceed some value R. Then the signals can be detected by examining the stars lying in a sphere of radius R around the Sun. How many stars will have to be examined in this way? The mean interstellar distance in the neighborhood of the Sun is about 2.2 pc (i. e., about 7 light years). The stellar density here is thus 0.1 stars per pc^3 or 0.003 stars per cubic light year. Let $R = 1000$ light years. A sphere of this radius will contain 10 million stars. The number of suitable candidates can be reduced if we remember that only a small fraction (at most 1%) may have planetary systems capable of sustaining life. We are thus faced with a very difficult and highly undeterminate problem: to choose a few hundred thousand stars from among 10 million which are likely to sustain highly developed civilizations. Ironically, the situation is much simpler with the search for civilizations in other galaxies (this problem has been treated above). Let us return to stars, however.

An optimum system for a search for signals sent from an unknown direction comprises directional fixed antennas whose beams cover the entire sky. If the sending EC transmits in all directions in space, it can be detected without difficulty. We have seen, however (Table 3.5), that this isotropic transmission requires a tremendous transmitter power and a highly directional receiving antenna. For a distance of 1000 light years and a transmitter band width of 1000 MHz, a power of the order of 10^{24} watt is required (this power is available only to type II civilizations) and the receiving antenna should have an effective area of 10^4 m^2 (assuming $T_n = 10°K$). This antenna has a directivity of 10^8 in the centimeter range, and some 100 million such antennas will be needed to cover the entire sky. Such a detection system clearly may be set up within the next 100 years. However, this project falls beyond the current financial resources of mankind.

The requirements regarding antenna area, the number of receiving antennas, and transmitter power can be relaxed if a directional transmitting antenna is used. Following V. A. Kotel'nikov, let us consider two civilizations A and B, distant R from each other. Civilization A transmits with a highly beamed antenna, and civilization B is at the

receiving end. Civilization A is not aware of the location of civilization B and the direction in which the signals should be sent is not known to start with. The antenna beam should therefore "trace" the entire sky. Let τ be the transmission length and ω the antenna solid angle $\left(\omega = \frac{4\pi}{g}\right)$. To scan the entire celestial sphere, it takes

$$t_0 = \frac{4\pi}{\omega}\tau = g\tau. \tag{3.83}$$

Suppose that civilization B has a detection system which comprises an assembly of directional antennas covering the celestial sphere. One of these antennas is aimed at civilization A. The receiver connected to this antenna records a signal at the time when the transmitting antenna of civilization A is aimed at civilization B. The signal detection experiment will take a time t much longer than t_0. The signal from civilization A will therefore be picked up several times, at equal time intervals t_0. In this way, it can be reliably distinguished from random noise. The time of detection can be somewhat reduced if civilization A, instead of scanning the entire sky, will concentrate on a limited number of suitable stars lying in a sphere of radius R (which naturally includes civilization B) and then send signals only in the direction of these chosen stars, shifting the antenna from one star to another.

Suppose that civilization A is distant 1000 light years, the transmitter power is 10^{17} watt (this is available to type I civilizations), $\Delta f = 1000$ MHz, transmitting antenna directivity $g = 10^9$, $\lambda = 3.5$ cm, transmission time $\tau = 3$ sec. From (3.83) we then find that $3 \cdot 10^9$ sec or 100 years will be needed to scan the entire celestial sphere. The time to scan all the stars inside a sphere of 1000 light years radius is $3 \cdot 10^7$ sec = 1 year. If only the most suitable stars are scanned (assuming that about 1% of the stars will support advanced civilizations), the total scanning time will be $3 \cdot 10^5$ sec or about 3.5 days. With a transmitter power of 10^{17} watt and a band width of 1000 MHz, a fairly humble receiving antenna with an effective area of about 100 m^2 (and $T_n = 10°K$) is required to detect signals over a distance of 1000 light years. One million such antennas will cover the entire sky simultaneously. Kotel'nikov proposed using multibeam antennas (technically, this is feasible, since each antenna is stationary), and the number of covering antennas can be reduced at least by one order of magnitude in this way. To reduce the number of antennas even further, he suggests dividing the sky into several areas and studying each area separately. This naturally will lengthen the detection time. Thus, in our case, when the scanning time for all the suitable stars is about 3.5 days, the celestial sphere can be divided into 10 parts, scanning each area in 36 days (during this period, the signal should appear at least ten times); the entire experiment will then be completed in 1 year, and it will require 10^4 ten-beam antennas. If $R = 100$ light years, the scanning time for the suitable stars is 300 sec. The sky can be divided into 10^4 areas and, scanning each area for 3000 sec (the signal will appear at least 10 times during this period), we will complete the experiment again in 1 year. The receiving antenna area needed to detect signals over a distance of 100 light years is 1 m^2, and the antenna directivity is 10^4, i.e., a single antenna will cover each of the areas! If $R = 10^4$ light years, the scanning time for

the most suitable stars will be $3 \cdot 10^8$ sec or about 10 years. Further division of the sky into areas will greatly prolong the experiment. And yet, for complete coverage of the sky at this distance, we will need 10^8 antennas of 10^4 m^2 each. Thus, given the transmitter power and bandwidth, the directivity of the transmitting antenna, and the transmission time τ, we can establish the optimum distance to other civilizations for which multiantenna detection systems are practicable. In our example, this optimum distance is of the order of 1000 light years.

The above reasoning applies to the search for wide-band signals, as well as monochromatic signals. The only difference is that in a search for monochromatic signals each antenna of the detection system should be provided with a multichannel receiver. For example, consider Kotel'nikov's receiver with 10^9 channels, $T_n = 30°K$, $\Delta f = 0.3$ Hz; let the transmitting civilization use a 10^9 watt transmitter and a 10^5 m^2 antenna, sending 10 cm monochromatic signals of duration $\tau = 3$ sec. Signal detection will then require a multiantenna system whose parameters are listed in Table 3.7.

TABLE 3.7. The parameters of a detection system for monochromatic IC signals ($P_1 = 10^9$ watt, $S_1 = 10^5$m^2, $\lambda = 10$ cm, $\tau = 3$ sec, $N = 10^9$, $\Delta f = 0.3$ Hz, $T_n = 30°K$) according to Kotel'nikov /16/

Distance in light years	Number of stars in a sphere of radius R	Scanning time for all the stars	Scanning time for the suitable stars (1% of the total)	Receiving antenna area, m^2	Number of receiving channels for the entire sky	Number of areas into which the sky can be divided	Number of receiving channels for each area
2000	10^8	10 years	36 days	400	480 000	1	480 000
1000	10^7	1 year	4 days	100	120 000	10	12 000
500	10^6	36 days	9 hours	25	30 000	100	300
200	10^5	4 days	1 hour	4	4 800	1000	5

On the basis of these data, Kotel'nikov came to the conclusion that radio signals from civilizations of our (or slightly higher) level definitely can be detected if there is at least one such civilization in 10^6 stars. If there is only one civilization in 10^7 stars, its detection presents a much more difficult problem, but it is nevertheless feasible under certain conditions. One civilization in 10^8 stars is extremely difficult to detect by the present-day means.

Wide-band signals. Sky surveys

Consider the search for wide-band signals. When the band width is of the order of magnitude of the transmitted frequency, the artificial signal is similar to the radio emission of natural sources. This leads to two conclusions. First, wide-band signals can be detected using

conventional radio astronomical equipment. Second, to detect wide-band signals, we should first establish how to distinguish the artificial from natural signals. After all, before attempting to decode the signal, we must be sure that we are dealing with an artificial source, which has to be identified among a multitude of natural radio sources. This brings us back to the problem of artificiality criteria, discussed in §3.

Any systematic search for artificial sources should include as a first step the discovery of all the radio sources followed by sifting in accordance with the likely artificiality criteria. Complete sky surveys in the radio spectrum should thus be launched. The meter and the decimeter wavelength range has been studied in fairly great detail. Detailed catalogues have been assembled for these wavelengths, listing all sources with radio fluxes down to 10^{-26} watt/m^2 · Hz. The situation is much worse in the centimeter range, however. No complete sky survey has been carried out in this range, and yet it is at these wavelengths that the civilizations are likely to communicate. Therefore, one of the immediate tasks is the organization of a detailed sky survey in the centimeter range using high-sensitivity astronomical equipment.

What are the requirements to be met by a radio telescope used in this survey? We should naturally strive to minimize the total survey time. And yet the radio telescope should have a maximum sensitivity or, in other words, the receiving antenna should be made as large as possible. Sky surveys can conveniently utilize the diurnal rotation of the earth. Consider a radio telescope with the antenna axis fixed in the meridional plane. The diurnal rotation of the Earth will successively aim the antenna pattern at different areas of the celestial sphere, all lying on the same diurnal parallel. In 24 hours, the telescope will survey a ring strip of the sky of width $2\theta_h$, where $2\theta_h$ is the vertical width of the antenna pattern between half-power points. Now we can displace the antenna through a distance $2\theta_h$ in declination, and it will survey a new annular strip during the next day; this strip adjoins the previous one and has the same width $2\theta_h$. The total time to survey the entire sky will clearly be

$$T_0 = \frac{\pi}{2\theta_h} \approx \frac{\pi h}{\lambda} \text{ days}, \tag{3.84}$$

where h is the vertical dimension of the radio telescope dish. The survey time thus decreases as θ_h increases or as the vertical dimension h of the dish decreases. For a given surface area, the minimum survey time is ensured if the vertical dimension of the reflecting surface is much less than the horizontal dimension, while the vertical dimension of the antenna pattern is much greater than the horizontal dimension ($\theta_h \gg \theta_l$). In other words, the radio telescope should have a "knife-edge" antenna.

Consider a radio telescope with an antenna in the form of a paraboloid of revolution 50 m in diameter (surface area 2000 m^2). This antenna has a symmetric pattern, whose width at 1 cm wavelength is $2\theta_h = 2\theta_l = 2\theta_{0.5} = 2 \cdot 10^{-4}$ radian = 40". Inserting this value of θ_h in (3.84), we obtain for the total survey time $T_0 = 43$ years. Let us now consider an antenna in the form of a parabolic cylinder with horizontal span $l = 400$ m and height $h = 5$ m.

This antenna, for the same geometrical area of 2000 m^2, has a "knife-edge" pattern with $2\theta_l = 5''$ and $2\theta_h = 7'$ (at 1 wavelength), and the total survey time will be 4.3 years. We see from this example that a radio telescope with a "knife-edge" antenna pattern not only greatly reduces the survey time, but also ensures a high resolving power (at least in one coordinate). When considering very large radio telescopes, whose size approaches the limit fixed by effects associated with radio brightness fluctuations of the metagalactic background and the atmosphere, we notice another important advantage of "knife-edge" antennas: they ensure the maximum sensitivity for a given antenna surface.

Two "knife-edge" antenna designs are currently known: Kaidanovskii and Khaikin's variable profile antenna (VPA) and the Krauss radio telescope. Figure 48 is a photograph of the large Pulkovo radio telescope with a variable profile antenna. The telescope is made up of separate shields, mounted along the arc of a circle. Each shield can be moved along the circle radius, turning in azimuth and position angle. By appropriately moving the separate shields, the reflecting surface can be rearranged so that the radio telescope is aimed at a desired point of the sky. The horizontal width of the antenna pattern is determined by the horizontal span of the antenna (the length of the chord spanned by the working sector); the vertical width for observations near the horizon is determined by the height of the shields. As the position angle increases, the vertical width of the antenna pattern diminishes, and in the zenith (when the VPA is a closed circle) it is equal to the horizontal width: the "knife-edge" pattern is thus transformed into a "pencil-beam" pattern. This effect increases the survey time. Krauss' radio telescope is more suitable for sky survey purposes (Figure 49). It consists of two separate reflecting surfaces: a fixed parabolic reflector whose optical axis is aligned in the meridional plane, and a moving plane reflector which may be rotated about a horizontal axis, ensuring observations at various position angles in the meridian. This radio telescope has a "knife-edge" pattern, whose vertical width is determined by the height of the parabolic reflector and is independent of the source position angle. A slightly modified form of this radio telescope, operating at 21 cm, has been recently built in France (Figure 50).

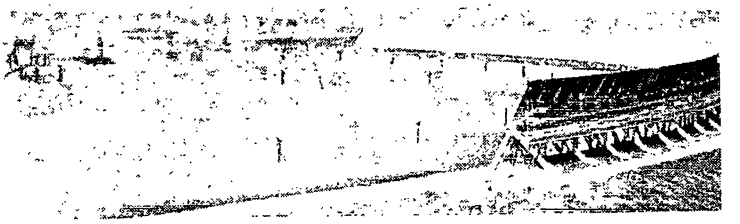


FIGURE 48. The large Pulkovo radio telescope with variable profile antenna.

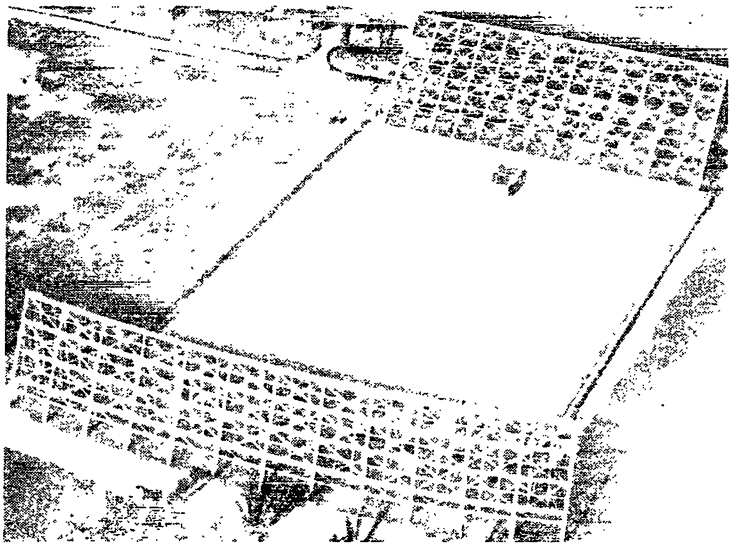


FIGURE 49. The Kraass radio telescope (USA).

When the sky survey has been completed, the sources should be sorted out according to some criteria. The criterion of angular dimensions is probably the most suitable to this end. One of the possible approaches is to identify all the sources with angular dimensions less than 0".1 and radio fluxes up to 10^{-27} watt/m^2 · Hz (omitting all the natural sources of large angular dimensions). This problem can be solved with high-sensitivity radio interferometers consisting of large antennas of $10^3 - 10^4$ m^2 areas separated to a distance of the order of $10^6 - 10^7$ wavelengths (in the centimeter range).

The selection of radio sources with angular dimensions of less than 0".1 should be regarded as the first preliminary stage of the program. Radio interferometers with an ultralong base, using the existing network of radio telescopes (a global system of radio interferometers), will attain a resolving power of 0".001 (a resolving power of 0".005 has already been attained). In the future, Earth spacecraft radio interferometers will probably be created. This will ensure bases of the order of 1 a.u. and reach resolving powers of the order of 10^{-8} angular second in the centimeter wavelength range. The selected sources will then have to be carefully studied using the various artificiality criteria. This opens wide horizons for future studies.

From the point of view of radio astronomy, artificial radio sources must possess certain unusual properties, i.e., an artificial source is a priori a peculiar radio source. The problem of discovering and studying peculiar radio sources is one of the basic tasks of radio astronomy. In this respect, our problem of search for extraterrestrial civilizations is closely linked with one of the most topical and pressing problems of radio astronomy.

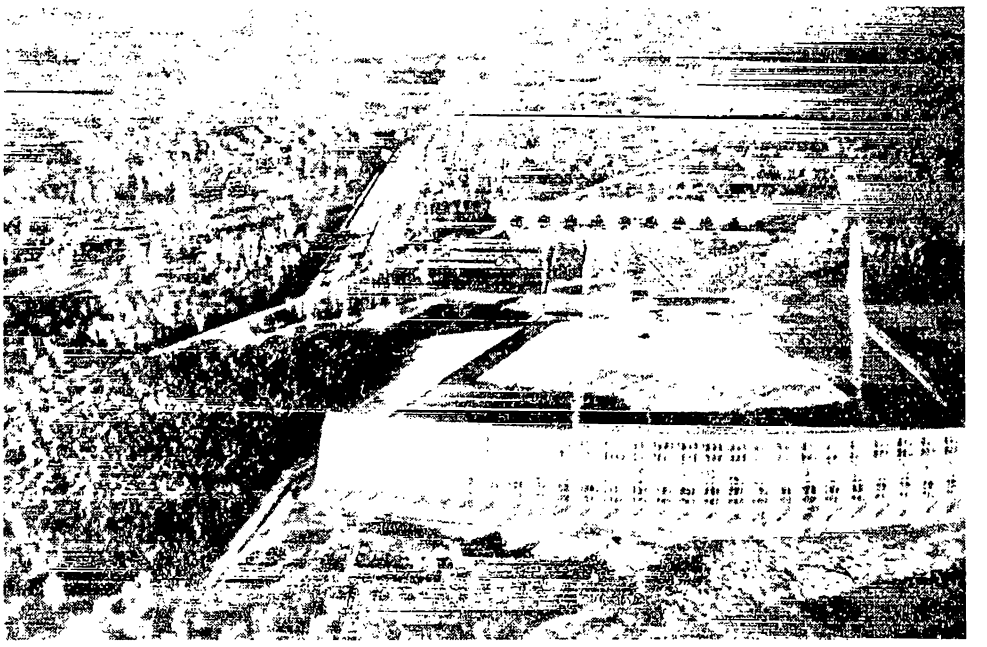


FIGURE 50. The Nançay radio telescope (France), operating at 21 cm wavelength.

The effective antenna area is 7000 m^2, horizontal beam width 3'.5, vertical beam width 20'. The receiver has 15 channels of 280 kHz band width each.

Bibliography

1. Cocconi, G. and P. Morrison. — Nature, Vol. 184: 844. 1959.
2. Cameron, A. (Editor). Interstellar Communication. — New York. Benjamin. 1963.
3. Vnezemnye tsivilizatsii (Extraterrestrial Civilizations). Proceedings of a Conference, Byurakan, 20—23 May, 1964. — Izd. AN Arm. SSR. 1965.*
4. Hartley, L. V. L. Transmission of Information. — BSTJ, 7 (3): 535—563. 1928.
5. Shannon, C. E. Communication in the Presence of Noise. — PIRE, 37 (1): 10—21. 1949.
6. Troitskii, V. S. Nekotorye soobrazheniya o poiskakh razumnykh signalov iz Vselennoi (Some Considerations on the Search for Intelligent Signals from Space). — In: Extraterrestrial Civilizations,* /3/, pp. 62—71.
7. Webb, J. Discovery of Intelligent Signals from Outer Space. — In: Interstellar Communication /2/.

* [See footnote on p. 11.]

8. Siforov, V.I. Nekotorye voprosy poiska i analiza radioizlachenii ot drugikh tsivilizatsii (Some Aspects of the Search for Radio Signals from other Civilizations and their Analysis). – In: Extraterrestrial Civilizations* /3/, pp. 78–83.
9. Shklovskii, I.S. Izluchenie "misteriuma" kak lazernyi effekt ("Mysterium" Radiation as a Laser Effect). – Astr. Tsirk., No. No. 372 : 1–8. 1966.
10. Kardashev, N.S. Peredacha informatsii vnezemnymi tsivilizatsiyami (Information Transmission by Extraterrestrial Civilizations). – Astron. Zhurnal, Vol. 41 : 282. 1964.
11. Slysh, V.I. Radioastronomicheskie kriterii iskusstvennosti radioistochnikov (Radio-astronomic Artificiality Criteria of Radio Sources). – In: Extraterrestrial Civilizations* /3/, pp. 38–42.
12. Gudzenko, L.I. and B.N. Panovkin. K voprosu o prieme signalov vnezemnoi tsivilizatsii (Reception of Signals Transmitted by Extraterrestrial Civilizations). – In: Extraterrestrial Civilizations* /3/, pp. 43–45.
13. Sholomitskii, G.B. Fluktuatsii potoka CTA-102 na volne 32,5 cm (Flux Fluctuations of CTA-102 at 32.5 cm wavelength). – Astron. Zhurnal, Vol. 42 : 673. 1965.
14. Golei, M. Coherence of Intelligent Signals. – In: Interstellar Communication /2/.
15. Shklovskii, I.S. Vselennaya, zhizn', razum (Life and Intelligence in the Universe). 2nd Ed. – "Nauka." 1965.
16. Kotel'nikov, V.A. Svyaz' s vnezemnymi tsivilizatsiyami v radiodiapazone (Radio Communication with Extraterrestrial Civilizations). – In: Extraterrestrial Civilizations* /3/, pp. 72–77.

* [See footnote on p. 11.]

Chapter IV

METHODS OF MESSAGE DECODING

§1. INTRODUCTION

The problem of signal decoding evidently occupies an important position among the various topics relating to communication with interstellar civilizations.

Every astronomer, analyzing the signals from various celestial objects, uses his own decoding system in the interpretation of his observations. However, the information discussed in connection with extraterrestrial civilizations is not the kind of information confined to the particular source: this information in principle reflects the structure of the Universe, including the organization of a certain society of "intelligent beings," i.e., it covers approximately the same scope as the "terrestrial" literature.

A characteristic feature of the problem of decoding of messages from extraterrestrial civilizations is the virtually total lack of any prior information or knowledge about these civilizations. We are thus faced essentially with a problem of decoding an arbitrary text.

Until recently, the problem of decoding of arbitrary texts did not attract particular attention in linguistics. Nevertheless, some decoding methods are available, using a minimum of preliminary information about the text. The general ideas underlying these decoding methods appear quite interesting, and the experimental results are very promising. It is hoped that the "extraterrestrial bias" will provide a strong stimulus to the development of this direction in linguistics.

There is always a chance that some accidental development will help to decode the message. It would seem that the messages from extraterrestrial civilizations would be organized in such a way as to simplify their decoding as far as possible. It is more prudent to assume, however, that the decoding of these messages will present considerable difficulties, no smaller, say, than the decoding of inscriptions in ancient lost languages. This approach is particularly important in that it prepares us for a linguistic struggle with extraterrestrial messages, and does not limit our task to mere detection. For a professional linguist, the tackling of codes and ciphers is a highly attractive occupation, which requires deep insight into the structure and the nature of language.

Interstellar linguistics also presents another problem (besides decoding). This is the problem of creating the most effective language for interstellar communication. It is particularly attractive in that every linguist goes all the way toward creating a certain consistent language, whereas there can hardly be a man capable of developing a full range of decoding methods.

However, the topical interest of this problem lies elsewhere, since interstellar communication cannot take the form of a dialogue. In the best case, a response to a message will be received after several centuries. If, on the other hand, extraterrestrial civilizations will take longer over responding to a message than it takes us to crack their code (or will lose interest altogether), interstellar communication will never progress beyond the realm of science fiction.

Interstellar communication is apparently not unlike literary activity: the messages are broadcast by the author civilization in all directions (just like books sent to various libraries and bookstores); the sender does not expect any response, just as the author never writes a book for the sake of a review. The reward is the privilege of getting acquainted with messages sent from other worlds.

Mankind will clearly make its first steps in interstellar society as a reader, rather than a writer. The problem of message decoding is therefore much more pressing than the problem of developing interstellar languages, at least at the present stage.

The aim of this chapter is to acquaint the reader with new linguistic methods of message decoding.

These methods are computer oriented and therefore basically reduce to algorithms, sets of instructions for a computer. For the reader's convenience, the algorithms are presented in generalized condensed form, with omission of most of the insignificant details.

The aim of decoding methods is two-fold. In practice, they are designed for cracking code messages. Theoretically, decoding algorithms present definitions of the linguistic features that they recognize in the message. In this respect, they are of particular interest to the professional linguist. The main significance of algorithms from this theoretical point of view is that they provide general methods of analysis, suitable for repeated application. The generality of the algorithms imposes natural restrictions on the intuition and the whim of the linguist.

This two-fold aim presents different requirements to be satisfied by the algorithms; on the one hand, they should provide accurate results, and on the other hand, they should be as free as possible from arbitrary features and logical ambiguities. For example, we tried to avoid the use of "empirical" numerical constants. In cases when the arbitrary approach was inevitable, we tried to apply simple solutions. This includes the construction of "estimate functions" of maximum simplicity. Occasionally, we reproduce algorithms which are known to provide unsatisfactory results, because their "scheme" may prove helpful in future work.

The reader will notice that the material presented in this chapter is of uniform interest. We wanted to focus our attention on the "basic" algorithms — the algorithm of identification of two groups of letters, the semantic algorithm, the algorithm of search for the sentence graph, algorithms identifying code sequences and morphemes, pattern decoding algorithms, and letter comparison algorithms.

Some readers may think that algorithms identifying vowels and consonants have only remote relation to interstellar communication problems. We want to stress, however, that all the algorithms are amenable to a more general interpretation (this point will be discussed in greater detail later on).

§2. THE CONCEPT OF A MESSAGE, ITS INTELLIGIBILITY AND MEANINGFULNESS

Definition of message

The aim of the present section is to provide an exact formulation of the basic concepts and problems encountered in decoding. It is generally assumed (and rightly so) that decoders deal with messages which should be understood and translated into a known language. When decoding messages received from outer space, there is an important preliminary stage: it is necessary to establish whether or not the message is intended for decoding (or is worth the effort). In other words, we have to establish first that the message is meaningful. These concepts of intelligibility and meaningfulness will be analyzed in this section.

A message is a system M of three sets: the alphabet (the set of letters) $A=\{a_i\}$, the set of positions $L=\{l_i\}$, and the text $T=\{a_i l_j\}$, or the product of the set of letters with the set of positions, i.e., the set of pairs of the form $a_i l_j$, where $a_i \in A$, $l_j \in L$.

In case of a general message, no restrictions are imposed on any of the three sets; they may be either finite or infinite, mathematically they may present groups, rings, spaces, etc. This is clearly a very general concept, and for many practical purposes the concept of a message should be properly restricted.

The set A is generally assumed to be finite or at least enumerable; a metric or topology is defined on the set of positions. Finally T, the set of text inclusions, is generally characterized as a one-to-many mapping of the set of letters into the set of positions, i.e., to each position l_j is assigned a single letter a_i, but any letter a_i can be found in any number of positions in the set L.

The last condition leads to the highly important concept of an "absolute frequency" of a letter a_i.

The absolute frequency of a letter a_i (of $\varphi(a_i)$) is defined as the power of the set $\{a_i l_x\}$, i.e., the set of all textual pairs containing the letter a_i.

The metric of the set of positions can be extended to the text: a textual distance between the pairs $a_i l_t$ and $a_k l_s$ is naturally defined as the distance between l_t and l_s.

Distance in the set of positions can be defined in a different way also. For example, we can define the relation of adjacency by specifying what pairs are adjacent and what are not.

Let $A=\{a_i\}$ be the Russian-language alphabet. L consists of two ring sections. The rings can be moved at random one relative to the other, and the text appears as shown in Figure 51.

In this case, we cannot define distance between positions on two different rings, but for each position we can identify two adjacent positions (on the same ring).

This definition of a message may look too general. Why not define a message in the usual way, as a string of letters?

There are examples, however, which make the conventional concepts look quite unnatural. A drawing may not be considered as a message; on

the other hand, a linear scan of the same drawing is a message. The definition of a text in terms of mapping into a graph suffers from similar shortcomings.*

Nevertheless, it is desirable to formulate less general definitions of a message for particular uses.

A message in a general sense will thus be characterized by the following additional features: 1. The sets A, L, and T are finite. 2. An adjacency relation v is defined on the set L, satisfying the following properties: a) if $l_i v l_j$, then $l_j v l_i$; b) for any position l_x, except two (l_p and l_h), there are two adjacent positions, i.e., there exist two positions l_y and $l_z (l_y \neq l_z)$ such that $l_x v l_y$, $l_x v l_z$, $(l_x \neq l_y, l_x \neq l_z)$; c) l_p and l_h have one adjacent position each; d) any partition of L into two parts generates adjacent positions which belong to different parts.

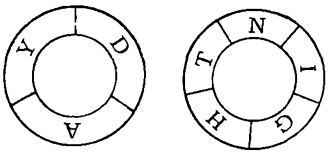


FIGURE 51. Example of an unconnected text.

A message is probably always expressed by some text. Cases when not all the letters of the alphabet occur in the text may give rise to some doubts. Anyhow, the concepts of "message" and "text" are largely interchangeable, and we will assume that they are synonymous.

We will now proceed with the problem of identifying what messages are worth decoding.

Artificial and natural messages

To distinguish the signals from "ordinary" stars and signals transmitted by intelligent beings, we speak of "natural" and "artificial" messages, respectively. It is often assumed that artificial signals from outer space should markedly differ from natural signals in some unusual property, which cannot be accounted for by physical considerations (see Chapters I and III).

However, many quite unexpected phenomena eventually find "natural" explanations; yet there are examples of artificial communications which can be made as close as desired to natural messages.

Consider the hypothetical case of a high-quality 3D cinema. If the screen is inaccessible, there is absolutely no way to distinguish between the view through a window (natural message) and the view projected on the screen. Note that the invention of holography will probably lead to development of three-dimensional movies with precisely these properties.

Another example is less tangible, but it has bearing on the case of signal search in outer space: consider a variable star observable at point A and not observable at point B. At the same time, point A is within the visibility range from point B. An observer at A may inform an observer at B of the exact behavior of the star by constructing a model which exactly simulates the behavior of the variable star. If the quality of the model is sufficiently high, the signals from the model will be indistinguishable from signals emitted by the real star. Nevertheless, the signals from the model are artificial, and the signals from the star are real.

* A graph is a union of two sets: the set of "vertices" and the set of "sides," in two-to-one correspondence (i.e., each side joins two vertices). Graphs can be presented in graphical form: a typical example is an airline flight network, with towns acting as "vertices" and flights as "sides."

These examples illustrate the futility of all attempts to devise a general formal definition of the concept of artificiality to be applied as a general criterion of signal selection.

We will now try to show that even legitimate artificial "meaningful" messages may have a form which will rule out all possibility of decoding. If there are signals of this kind, they will remain unintelligible despite their probable artificiality.

Any message can be "scrambled" in such a way that it will be understandable only to an observer with adequate "descrambling" knowledge, or in other words an observer who has in his possession the "key" to the cipher. In some cases, messages can be descrambled even if the key is not available to start with. However, if the key volume is comparable with the volume of the coded message, the text can be so scrambled as to become theoretically undecipherable by any conceivable technique. This observation is due to Shannon /8/. Examples of such scrambling techniques are easily constructed.

Consider a Russian-language text N letters long. The position of each letter in the text can be specified by its running number i from the beginning of the text. Each number i ($1 \leqslant i \leqslant N$) is written on a separate card and the cards are then shuffled and spread in a random sequence. In the resulting sequence C, the card i will occupy position j from the beginning. If the appropriate letters of the original message are substituted for these positions, we obtain a coded message. To decipher the message, we require the sequence C (the key). In deciphering, the j-th letter of the coded message should be moved to a position identified by the j-th element of the key.

If the key is not known, this message clearly cannot be decoded; the coded sequence of letters is truly a random sequence. Although the relative frequencies of the individual letters correspond to the frequencies of the Russian language, this fact can be easily concealed by adding as many rare letters to the text as is needed to equalize all the frequencies.

Intelligibility of a message

We are thus concerned not just with messages sent by intelligent beings, but with intelligible messages, i.e., messages that can be understood.

Are there specific criteria distinguishing intelligible messages from unintelligible ones?

Suppose that only part of the text is available for examination, i.e., the text has been partitioned into an accessible and an inaccessible part. If, by examining the accessible part of the message, we can predict what the inaccessible part probably contains, we will say that the message is intelligible relative to the given partition. If the inaccessible part can be predicted for any partition of the text into an accessible and an inaccessible part, we say that the message is completely intelligible.*

We will show in a few examples that this definition of intelligibility does not contradict the usual sense of this word.

Indeed, the sentence "Pushkin was borr in the 18th paragraph" is unintelligible because if the accessible part of the message is "Pushkin was born in the 18th..." it is impossible to predict that the next part of the text is

* Intelligibility relative to a particular partition is a numerical function of the partition. No formal expression for this function can be given at this stage, however.

"... paragraph." One would naturally expect a word (or a group of words) signifying a period of time (e.g., "... century").

The sentence "Pushkin was born in the 18th centuries" is equally unintelligible, since we expect the word "century" in singular and not in plural. The sentence "Pushkin was born in the 18th siècle" is again unintelligible, since there is no reason to expect a French word in an English sentence. (If the sentence is unintelligible, but it is clear how it should be modified to make it intelligible, we generally say that the sentence is incorrect.)

A picture of a man with the left leg replacing the right arm is unintelligible; an object which looks like a log but sinks in water behaves unintelligibly; a random sequence of letters is completely unintelligible.

Let us now consider examples of intelligible messages. An infinite sequence of letters "... aaa..." is intelligible relative to any partition, since the only reasonable prediction is "the i-th position of the unexamined part of the sequence is occupied by the letter a," and this prediction is always true.

A message of the form "... abcabcabc..." is intelligible relative to any partition for which the accessible part is long enough to reveal the three-letter cycle. The picture of an infinite straight line is completely intelligible. The sentence "Pushkin was born in the 18th century" is intelligible to an educated English-speaking person, i.e., a person capable of predicting the sequence of occurrence of words in English-language sentences. The picture of a man is intelligible to all intelligent beings who have seen a man alive or in other pictures. Any periodic process is intelligible relative to partitions revealing a sufficiently long part of the message.

We will now show that the ability to predict is based on knowledge of certain special properties of the text or its components. Consider the sequence of words "Napoleon invaded Russia in ..." It can be completed to read "Napoleon invaded Russia in 1812," but an equally intelligible sentence will be "Napoleon invaded Russia in the 19th century." Formally and morphologically the words "...1812" and "... the 19th century" are as far apart as the expressions "18th century" and "18th paragraph" in the previous example. And yet there is a conceptual similarity between these expressions, i.e., they fall in the class of words which are "close in meaning" or, to use a different phrase, their "semantic distance" is small.

Texts may comprise small elements (e.g., words), as well as large elements (e.g., sentences). If we know what typical word combinations make a sentence, we can predict the missing words having read through a part of a sentence (typical examples are combinations of so-called grammatical classes, e.g., the "nominative case," "finite verb," etc.). Correct prediction thus requires breaking the text into sentence-like parts.

This partition may be quite complex; compound sentences are a common occurrence in modern languages. We should therefore try to assess the closeness of "words" not in terms of their "adjacency," but by some other method.

Meaningfulness of a message,
predictive system, language

The information required for effective prediction of textual elements can be indicated by an appropriate re-coding of the message, whereby semantically close parts are written in one common form, and the semantically dissimilar parts are written in different form; textual elements combined into larger components should be enclosed in brackets; "semantically close" parts should also be textually close. This re-coding and rearrangement of the text will be called interpretation.

The best interpretation is clearly that which ensures the highest intelligibility. The selection of the best interpretation may be regarded as message decoding in the narrow sense of the word or as partial decoding. The correspondence (mapping) between the elements of the message and the best interpretation will be called the "predictive system" of the message or its complete grammar. The predictive system, on the one hand, is close to conventional grammars and, on the other, to dictionaries.

The language is naturally defined as the set of messages with the same predictive systems. In other words, the messages in one language are constructed "in the same way." If there is a correspondence between the elements of the best interpretations of two messages, a certain correspondence also can be established between the elements of the messages. In this case, one text is a translation of the other.

The translation of a coded message can be regarded as the ultimate aim of decoding. We will see in §8 that it is easier to look for correspondence between the elements of the messages than for correspondence between the elements of the best interpretations. For decoding purposes, we should therefore study the predictive systems of known, as well as unknown, languages.

The above examples of intelligible messages are disappointing to a certain degree. Intelligibility clearly does not exhaust all the properties of messages which have bearing on successful decoding. We will try to make use of the fact that interstellar messages are probably constructed in a special way so as to facilitate decoding to a maximum degree.

The best interpretation in this case should be easily identifiable, i.e., it should be readily distinguishable from the other interpretations. If the quality of an interpretation is assessed in terms of its intelligibility, the identifiability of the best interpretation can be defined as the difference between the intelligibility of the best interpretation and some other (e.g., worst) interpretation. The identifiability of the best interpretation is a fundamental property of messages intended for decoding; it is this property that we call meaningfulness.

It is readily seen that messages without sufficiently intelligible interpretations are not very meaningful; on the other hand, messages for which all the interpretations are intelligible are not very meaningful either. This accounts for the triviality of the examples described on p. 138: no low-intelligibility interpretations can be constructed from these examples. Note that messages expressed in normal languages (without any coding) are "intended for deciphering" in a certain sense, and are therefore highly meaningful.

Let us briefly consider the concept of "external meaningfulness." Consider two partitions R_i and R_j of a text T into an accessible and an inaccessible parts, $R_i = T_i^{acc}, T_i^{inacc}$; $R_j = T_j^{acc}, T_j^{inacc}$, $T_i^{acc} \subset T_j^{acc}$ (the symbol T_i^{acc} identifies the accessible part of the message, the symbol T_i^{inacc} the inaccessible part). To each of these partitions corresponds a certain value of the intelligibility Π in the best and the worst interpretations, $\Pi(R_i)^{best}$, $\Pi(R_i)^{worst}$, $\Pi(R_j)^{best}$, $\Pi(R_j)^{worst}$. The increment $\Delta\rho$ of meaningfulness on passing from partition R_i to R_j is expressed in the form

$$\Delta\rho = [\Pi(R_j)^{best} - \Pi(R_j)^{worst}] - [\Pi(R_i)^{best} - \Pi(R_i)^{worst}].$$

The value of $\Delta\rho$ can be defined as the external meaningfulness of a message whose text is $T_j^{acc} \setminus T_i^{acc}$ (i.e., accessible in R_j and inaccessible in R_i). In particular, if T is the text of the message about the outside world provided by the sensory organs during the entire span of human life, $\Delta\rho$ is the meaningfulness increment acquired as a result of a message with the text $T_j^{acc} \setminus T_i^{acc}$.

A particular example of the application of these principles will be described in §6.

§3. TRADITIONAL METHODS OF MILITARY AND LINGUISTIC DECIPHERING

Military deciphering

Deciphering of coded messages is a common practice in two fields of human activity: it is often the task of historians and linguists (in their attempts to read texts in lost languages), and also of military and diplomatic personnel, who have to deal with intentionally coded messages in known living languages.

According to the literature (see, e.g., /14/), military deciphering techniques assume certain limited traditional forms (although, as we have seen, messages can be scrambled beyond all ability to decipher them).

A military cipher is difficult, and sometimes even impossible, to break. These ciphers, however, are fundamentally simple compared to the predicate system (grammar) of a real language. Coding is generally done through juggling with letter sequences which do not have any semantic relation to the actual text.

Let us consider some of the common ciphers /14/.

Simple substitution cipher. Each letter is replaced with an alternative symbol (generally another letter).

Transposition with a fixed period t. The entire message is divided into segments t letters long, and the same substitution is applied to each segment.

The Vigenère cipher and its modifications. The key is a sequence of t letters. It is written consecutively, as many times as is

needed, under the original message, and the two sequences are added modulo n, where n is the number of letters in the message alphabet.*

For example:

original message — LETTERNOTYETRECEIVED
key — TROYTROYTROYTROYTROY
cipher — EVHRXIBMMPSRKVQCBSMR

If the key is a single letter, the result is known as the Caesar cipher; coding can also be done using an aperiodic letter sequence (which produces an indecipherable cipher). In another cipher, each letter is replaced with a sequence of t symbols. In so-called "code systems," words, groups of words, or syllables are replaced with various letter combinations.

Deciphering is based on two fundamentally different approaches: the statistical method and the method of characteristic words. In the statistical method, the frequencies of the letters in the cipher are compared with the frequencies of the letters in the real language in which the message is presumably written (the real language statistics is obtained from a sufficiently large representative sample). If the frequencies of the letters in the language are close to the frequencies of some cipher elements, these elements are interpreted as the images of the corresponding letters.

In the method of characteristic words, we search for smaller component elements which repeat like the letters of certain characteristic words which are presumably contained in the cipher. These principles are used in the algorithms of §8.

In certain cases, all the possible ciphers of a certain class can be examined, and the text being analyzed can be applied to verify that a particular cipher has been used in that case. For example, if it is known that the Caesar cipher has been used, the probability of a particular cipher is a function of the intercepted cryptogram volume, and this function can be calculated.

Suppose that we have intercepted a cryptogram containing the part of a sentence "... creases to..." (where "creases" is the end of the word "increases"), coded in the Caesar cipher. If only the cipher of a single letter "c" has been received, the deciphered result may be represented by any letter of the English alphabet. In this case, the probability of each deciphering is equal to the probability of the corresponding letter.

If two letters (cr) are received, there are 26 different decipherings of the message (assuming that the Caesar cipher has been used). The probability of each version is equal to the probability of the corresponding pair of letters in the English language, and so on.

Table 4.1 (due to Shannon) lists the results of these calculations for sequences of up to five letters. Suppose that the enciphering was done by using the letter a_i. The letter sequences under the heading "Deciphering"

* Addition of letters modulo n is carried out as follows: let $i(a)$, $i(b)$, and $i(c)$ be the current numbers of the respective letters in the alphabet. Then c is obtained from the equality $i(a)+i(b)=i(c)$ if $i(a)+i(b)\leqslant n$, and $i(a)+i(b)-n=i(c)$ if $i(a)+i(b)>n$.

The cipher is deciphered with the aid of subtraction modulo n, i.e., $i(a)$ is found as follows: $i(a)=i(c)-i(b)$ if $i(c)-i(b)\geqslant 0$, and $i(a)=i(c)-i(b)+n$ if $i(c)-i(b)<0$.

TABLE 4.1

Deciphering	$N=1$	$N=2$	$N=3$	$N=4$	$N=5$
CREAS	0.028	0.0377	0.1111	0.3673	1.0000
DSFBT	0.038	0.0314			
ETGCU	0.131	0.0881			
FUNDV	0.029	0.0189			
GVIEW	0.020				
HWJFX	0.053	0.0063			
IXKGY	0.063	0.0126			
JYLHZ	0.001				
KZMIA	0.004				
LANJB	0.034	0.1321	0.2500		
MBOKC	0.025		0.0222		
NCPLD	0.071	0.1195			
ODQME	0.080	0.0377			
PERNF	0.020	0.0818	0.4389	0.6327	
QFSOG	0.001				
RGTPH	0.068	0.0126			
SHUQI	0.061	0.0881	0.0056		
TIVRJ	0.105	0.2830	0.1667		
UJWSK	0.025				
VKXTL	0.009				
WLYUM	0.015		0.0056		
XMZVN	0.002				
YNAWO	0.020				
ZOBXP	0.001				
APCYQ	0.082	0.0503			
BQDZR	0.014				
H (decimal units)	1.2425	0.9686	0.6034	0.2850	0

are the sequences obtained by subtracting from the cryptogram the sequences

$$
\begin{array}{ccccccc}
a_i, & a_i, & a_i, & a_i, & \ldots, & & \\
a_{i-1}, & a_{i-1}, & a_{i-1}, & a_{i-1}, & a_{i-1}, & \ldots, & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
a_1, & a_1, & a_1, & a_1, & a_1, & \ldots, & \\
a_n, & a_n, & a_n, & a_n, & a_n, & \ldots, & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
a_{i+1}, & a_{i+1}, & a_{i+1}, & a_{n+1}, & a_{i+1}, & \ldots &
\end{array}
$$

The first column of numbers ($N = 1$) gives the probability of single-letter sequences in the English language. These sequences provide the deciphering probability of a single letter "c". The second column contains the deciphering probability of the first two letters (cr), i.e., cr, ds, et, etc. The column $N = 5$ contains the deciphering probability of all the five letters of the text creas, i.e., creas, dsfbt, etc. In this column, the probability of the sequence creas is close to 1, and the other probabilities are close to zero.

Vacant spaces in the table correspond to very low probabilities. The probabilities were calculated from data about the frequencies of two- and three-letter sequences given in /8/.

The row H gives the entropy of the probability distributions for the five cases. The entropy H $\left(H = -\sum_i p_i \log p_i\right)$ was calculated from the values of p in this table, using decimal logarithms.

Linguistic deciphering

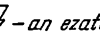
The deciphering of old texts is apparently more relevant for the purposes of interstellar linguistics. Students of old languages are forced to reconstruct their highly complex structures. Moreover, the texts are not scrambled intentionally, and they are therefore far from a random jumble of letters. The complexity of the natural languages, however, is responsible for the lack of a general deciphering method, despite the successful cracking of numerous old texts.

The various cases of successful deciphering of old texts are largely due to pure luck and to ingenious intuitive guesses, which will not work in other cases.

Thus, the world-famous deciphering of the Egyptian hieroglyphs is traceable to the discovery of bilingual inscriptions, i.e., the unreadable text was accompanied by its translation; the Hittite language was deciphered after a brilliant guess as to the nature of the related languages; the Creto-Mycenaean inscriptions were deciphered on the assumption (since proved correct) that the language in question was Greek.

We will quote here from the article by Hrozný (who deciphered the Hittite cuneiforms), describing the first breakthrough in his work. Note that the pronunciation of the individual cuneiforms was known at that time, and the meaning of ideograms — i.e., symbols representing concepts, and not sounds — was also familiar.

"The method of my work is best illustrated by considering the following sentence, one of the first whose meaning I was able to establish, and in which I recognized three Hittite words of Indo-European origin.* This cuneiform I read phonetically.**

nu  *-an ezateni wādar-ma ekuteni.*

"When I first came across this Hittite sentence, I knew only the meaning of the ideogram, which often, though not always, stands for "bread." Other parallels indicated that the suffix was accusative singular. Despite numerous other possibilities, it was reasonable to assume that a sentence dealing with bread will also contain the verb "to eat." I therefore started with the purely hypothetical assumption that the word "ezateni" signifies the concept of eating. Soon after that I noticed that the Hittite root "eza" stands for "to eat" in many other texts, and that another root with the same meaning is "ad," e.g., in the form "adanzi," they eat, which is probably identical with "eza." Then I compared, again purely hypothetically, these Hittite roots "ad," "ez" to the Latin "edo," the German "essen," etc. Other sources supplied me with an indication that "teni" is a second person plural ending in present and

* Related languages are languages arising from a common "source language." Words of close meaning in related languages have a similar pronunciation.

Russian, Ukrainian, Polish, Czech, Bulgarian, and Serbian are closely related languages (the so-called Slavic languages); more distant relations of Russian are German, Latin, Greek, and some Indian languages, forming together the so-called Indo-European family. Compare the following words:

Russian	мать
German	Mutter
Latin	mater
Greek	μητηρ
Sanskrit	mātar.

** I.e., its pronunciation was known. The cuneiform in the middle is an ideogram (a concept symbol). The phonetic composition of the equivalent word was not known.

future tense, so that I translated the first sentence as "you will eat bread." The next sentence looked parallel to the first: "vadar," a noun; "ma," a preposition; "ekuteni," a verb with "teni" ending. Since the word "vadar" was parallel to the word "bread," it probably also identified some simple food. The English word "water" and the Anglo-Saxon "watar" helped me to identify "vadar" as water.

"The noun "water" was thus followed by the verb "ekuteni," which correlated with the verb "ezateni," "you will eat." It therefore logically lended itself to translation as "you will drink." Later I found that besides the root "eku," to drink, there was also a close root "aku," to drink, e.g., in the word "akuvanna," to drink. The comparison of "akuvanna," to drink, with the Latin "aqua," water, was self-evident. I therefore translated the whole sentence as "you will eat bread and you will drink water."

It is clear from this excerpt that there could be no continuation to Hrozný's method: his guiding line was the assumption that the lost Hittite language was related to some known languages (words of similar meaning have close pronunciation in related languages); the rest of his arguments are fairly obscure, e.g., the contention that the two parts of the sentence are parallel and the frequent references to similar meaning that the same word has in other texts.

If the pronunciation of the letters is not known either, we have to rely on the occurrence of proper names in the text, as they are of international meaning to a certain extent; ideograms and hieroglyphs (whose meaning in a sense corresponds to that of a picture), pictures occurring in the text, or objects carrying inscriptions are very helpful in disclosing hidden meaning. There is generally some information available about the corresponding historical epoch, the wars which took place at the time of writing of the texts, the identity of rulers and leaders. Whole dictionaries are sometimes available (the Tangut and Mayan inscriptions). The decoding of old texts is thus largely dependent on the resourcefulness and the intuition of the linguist, who draws upon a tremendous treasure of information that may prove useful; all this, however, does not provide us with a set of general linguistic tools and techniques for text deciphering. Deciphering is closer to a one-time art, not quite understandable to the outsider, than to a practical science.

§4. SEQUENCE OF APPLICATION AND STRUCTURE OF DECODING ALGORITHMS

Sequence of algorithm application. Levels

In decoding a particular extraterrestrial message, we shall naturally have to lean heavily on our intuition and more or less incidental information. However, in so far as no extraterrestrial messages have been received, there is only one way for us to prepare for the future decoding, and this is by developing general decoding methods which will answer the greatest variety of needs.

In our opinion, these methods will be valuable only if they admit of clear-cut, unambiguous formulation. This condition is met by algorithms —

precise instructions for textual analysis which are so clear and comprehensible that a computer can carry them out. The computer-oriented approach is particularly helpful since deciphering involves processing of large blocks of information, which often cannot be done manually.

Given a complete system of such algorithms, we can visualize the operation as follows: the text to be decoded is fed into the computer, which then proceeds to translate it into one of the known languages. This ideal situation, however, is not very realistic.

There is, moreover, no need to carry the algorithm system to this extreme: it is sufficient to ensure algorithm solution of some "key" problems of decoding. This will leave relatively simple problems to be tackled by human ingenuity and intuition, the two properties presently unattainable by computers.

As we have noted before, decoding is primarily an activity intended to identify the "predictive system." Its second aim is to translate the original into one of the known languages. If there had been powerful decoding techniques meeting the second aim, we would not have had to search for the "predictive system" of the original.

Practical experience shows, however, that it is nevertheless better to search first for the predictive system and then to proceed with the development of translation techniques. Experiments carried out using the algorithm on p. 179 are highly illustrative in this respect. Moreover, the possible "non-interpretability" of extraterrestrial messages should be taken into consideration (see conclusion).

We will consider some algorithms whose importance for the construction of the predictive system is self-evident.

Traditional linguistics uses two techniques to distinguish between linguistic phenomena: one of these techniques resorts to real images and patterns that various expressions invoke in the mind of the language user, and the other makes use of our inherent ability to differentiate between correct and incorrect expressions in a particular language.

For example, verbs are distinguishable from nouns because verbs generally express a certain action or process, whereas nouns are identified with objects or abstract concepts. On the other hand, morphologically, a [Russian] verb is identifiable by its characteristic "endings," such as "л", "ла", "ло" (endings of past tense masculine, feminine, and neutral).

Modern applied linguistics, with machine translation as one of its most active branches, uses mainly information belonging to the second category, i.e., advance knowledge of certain morphological signs of linguistic phenomena is presupposed.

In the decoding of extraterrestrial messages, we naturally cannot resort to real images or patterns or to morphological features of the written language.

In the construction of decoding algorithms, we should proceed from the basic and most general properties of the phenomena. It is here that the linguist's interests lie.

An efficient decoding algorithm essentially provides a definition of the phenomenon that it is supposed to recognize. More precisely, we could simply define a particular linguistic phenomenon by what emerges from an arbitrary text when a particular decoding algorithm is applied to it.

These definitions are attractive in that they are applicable to unknown languages (i.e., they are highly general), they are extremely lucid (and can

be implemented by a computer), and are practicable, i.e., they provide a tool for recognizing various linguistic effects. The importance of these algorithms may turn out to be quite independent of the linguistic decoding aims.

Let us consider in more detail the structure and the sequence of application of decoding algorithms. Decoding algorithms clearly may use information disclosed by other decoding algorithms. If a certain algorithm B uses the ability to recognize a linguistic phenomenon defined by algorithm A, we will say that algorithm B is of a higher level than algorithm A. It would naturally be very unfortunate if algorithm A were at the same time of a higher level than algorithm B, since this would lead to a definition of an unknown in terms of another unknown. The only exception are algorithms which successively improve their own results. In this case, the seniority of the algorithms is determined by their seniority in the first iteration.

It is clear that there must exist a zero-level algorithm which does not use any information obtained by other decoding algorithms. Zero algorithms should differ according to the effect that the symbols have on the human sensory organs or on the decoding device. They should be associated with the minimum differences detectable by these sensory organs.

If we are developing algorithms intended for the analysis of written languages, the zero algorithms should naturally reconstruct the alphabet of the particular language by examining a certain sufficiently long text. The information required apparently reduces to the ability to distinguish between black and white squares, assuming that the text is covered with a very fine grille so that each cell is either black or white. The ability to identify the position of each cell is also required.

If spoken sounds are to be decoded, the zero algorithm should use the minimum acoustic differences. The variety of the zero algorithms evidently can be reduced by suitable conversion of signals with physical devices; e.g., speech can be represented by a chart plotting on paper the variation of air pressure.

At first glance, alphabet reconstruction is a very simple problem, which always can be solved after a brief inspection of the text. The human analyst is sometimes baffled by illegibility of the written text, but for machines the problem is complicated even for fairly clear inscriptions. Some insight into the problems involved in the identification of the phonetic alphabet may be gained by inspecting a segment of an oscillogram trace of Russian speech (Figure 52).

A curve representing a signal from outer space will be much less "legible": it will probably be distorted by strong noise. Curve manipulation is not among our strongest aptitudes, and it is therefore clear that the reconstruction of the alphabet of "elementary" signals will not be an easy undertaking.

The zero algorithm for written languages is thus expected to reconstruct letters as special combinations of dark and light squares. Although efficient algorithms for alphabet reconstruction can be developed in principle, no such algorithm is available at this stage. In §11 we will describe a rudimentary algorithm of this kind which is more of theoretical than practical importance.

Once the set of "elementary signals" has been identified, we can proceed with identification and analysis of larger elements. For languages close to human languages, the first level algorithms should be able to distinguish between various classes of letters of similar pronunciation, and also smallest

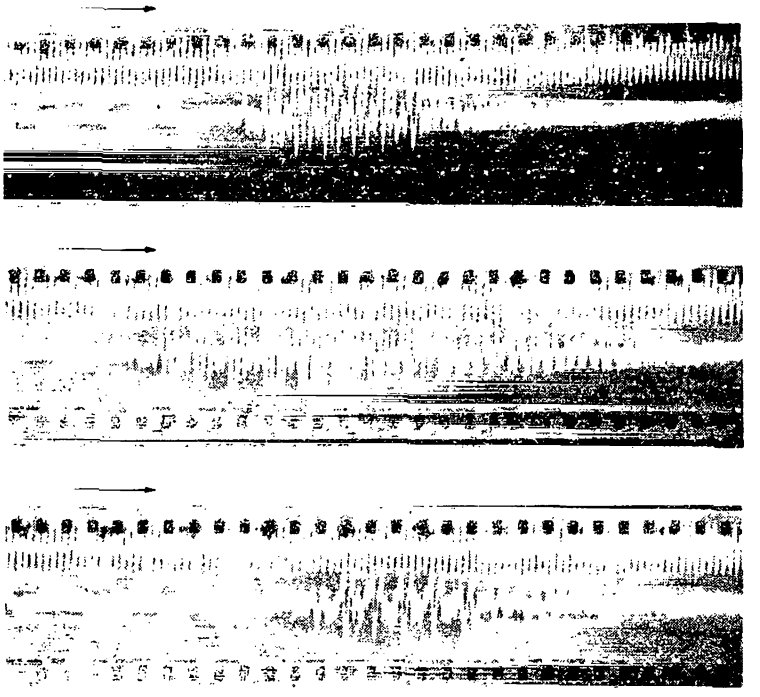


FIGURE 52. Oscillograms of Russian spoken syllables "tu," "ta," "pa":

All the syllables are stressed, extracted from a recording of individual sentences. The flat portion of the oscillograms corresponds to silence (closed mouth), then follows a burst which ensures the audibility of the sounds "t" and "p," and further large-amplitude fluctuations representing the vowels "u" and "a."

meaningful letter sequences which are not made up of smaller meaningful sequences (the so-called morphemes). This algorithm should be able to divide the message into morphemes even if no blanks are interposed between the words, since words are more complex elements than morphemes. In some orthographies, no division is made between words anyhow.

Second-level algorithms should locate the limits of the individual words and identify different classes of morphemes (such as semantically meaningful morphemes and auxiliary morphemes used as suffixes, prefixes, etc.). The third-level algorithms should search for classes of words and identify the limits of sentences. Higher-level algorithms should analyze those sentences and semantics.

Numerous algorithms of different levels may be quite similar. In some cases they are actually identical, differing only in the input material. Thus, algorithms identifying groups of letters with similar pronunciation can be used without any modification to identify different classes of morphemes; sentence-identifying algorithms are very similar to syllabilization algorithms; algorithms splitting the text into morphemes are not unlike the letter-identifying algorithms, etc.

In our discussion of the particular algorithms, we will always indicate on what different levels the particular algorithm may be used. However, essentially similar algorithms may each have its own specific features, generally related to the volume of processed information.

For example, an algorithm identifying classes of morphemes will provide an output which is about a hundred times larger than the output of the same algorithm operating in the letter-identifying mode. Because of these specific features, the programming of higher-level algorithms is substantially more complicated.

On the other hand, higher-level algorithms are naturally more interesting: they provide a fuller analysis of the text, permitting "long-range forecasting."

We will consider low-level algorithms, e.g., algorithms analyzing letter pronunciation. Not all of them are relevant for the decoding of extraterrestrial messages. However, they should all be considered as models: on a low level these algorithms solve problems which are much more topical and significant when tackled on a higher level.

There probably exists a limited number of different types of decoding algorithms, and we should therefore first examine one algorithm of each of the different types, before striving toward higher and higher levels.

Structure of algorithms: sets of alternatives, quality function, computation procedures. Types of algorithms

Various decoding algorithms have many features in common. We have indicated earlier that algorithms recognizing different linguistic phenomena in an unknown text may be used to provide the definition of the corresponding phenomena. However, somewhat more general definitions also can be offered. To this end, it suffices to formulate clearly the characteristic features of the linguistic phenomenon used in its identification. Computation procedure intended for recognition purposes (e.g., algorithms) may take different forms even for the same set of recognizable features. In our description of the decoding algorithms, we shall first describe the recognizable distinctive features, and then give the particular recognition procedure.

Recognition features in their turn fall into two categories: some do not require any computations or manipulations of the text, whereas others do. The former features are of binary character, i.e., they are present for a certain phenomenon and absent for other phenomena. Features of the second category express properties which are more prominent in this particular phenomenon than in other phenomena with the same "binary" recognition properties.

We say that features of the first group define the set of alternatives (the set of interpretations), whereas the features of the second group characterize the quality or the reliability of these alternatives. In other words, quality is a numerical function defined on the set of alternatives. The set of alternatives will also be called the set of permissible solutions, with a certain quality function.

We wish to emphasize one highly important property of quality functions. Until recently, linguists used definitions based on binary features (or, in general, features expressible by a finite number of digits). These definitions, however, proved to be quite complex: they contained numerous "exceptions" and were not particularly suitable for machine recognition.

This approach precluded the formulation of common definitions for similar phenomena in different languages. The concept of quality function greatly simplifies the "binary," i.e., logical, part of the definitions, and they acquire a greater generality. The reader will see that quality functions proved highly convenient in practice, since algorithms using these functions are generally programmed without much difficulty.

The aim of the recognition procedure (the algorithm) is to find a permissible solution which maximizes (sometimes minimizes) the quality function.

Whenever the set of permissible solutions is given and the quality function is defined, the determination of the permissible solution maximizing or minimizing the quality function becomes a purely mathematical problem.

Rigorous solution of mathematical problems of this kind to which decoding algorithms are reduced is mostly unknown. We tried to describe the most practical solutions, i.e., solutions which are sufficiently accurate to provide acceptable results and yet sufficiently simple to be implemented on existing computers.

Let us again consider the question of the various types of algorithms. The currently known algorithms intended for the analysis of predicate systems can be divided into the following groups:

Classification algorithms. These include the algorithms which divide the set of units being studied into nonintersecting subsets, e.g., the algorithms partitioning the set of words into classes which contain letters of similar pronunciation; algorithms partitioning the set of morphemes into classes of morphemes with identical "grammatical" properties (auxiliary morphemes vs. meaningful morphemes); algorithms identifying semantically close classes of words.

Matching algorithms. We use this term for algorithms which form small linguistic elements into larger linguistic units; e.g., the algorithm of morpheme identification, the algorithm of letter identification, the algorithm of sentence identification, and the algorithm of syllable identification.

Algorithms establishing semantic closeness. The visual closeness of words in a text does not always correspond to the actual semantic closeness of words. Similarly, in a linear scan of a two-dimensional pattern, adjacent elements are not the only close elements: elements separated by the length of one line are of course also close.

Algorithms of this kind include the algorithms which determine the so-called sentence graph (see p. 198). Note that knowledge of the "true closeness" of elements is essential for correct functioning of the matching algorithms.

Translation algorithms receive less attention in this chapter. Decoding apparently can be confined to algorithms compiling various bilingual dictionaries. In machine translation, algorithms synthesizing sentences in the product language are of considerable importance. In decoding, this problem can readily be left to the human operator.

The description of the various algorithms in this chapter does not correspond to the order indicated above. Simpler and more obvious algorithms, accompanied by examples, are given in §5 and §6, the others are deferred to §7 through §11.

TABLE 4.2

	й	т	д	т'	д'	н	н'	ц	с	з	с'	з'	п	б	п'	б'	б'	м	м'	ф	п	ф'	в'	ч	щ	ж	к	к'	д	х	е	е'	о	о'	а	а'	ы	ы'	ы	ы'	р	р'	л	л'
	j	t	d	t'	d'	n	n'	c	s	z	s'	z'	p	b	p'	b'	b'	m	m'	f	v	f'	v'	č	š	ž	k	k'	g	x	e	e'	o	o'	a	a'	ɨ	ɨ'	u	u'	r	r'	l	l'
Vocalic	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Consonantal	–	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+
Open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	–	–	–	–	–	–	+	+	+	+	0	0	0	0
Compact	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–	0	0	0	0	0	0	0	0
Low-level pitch	0	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	+	+	–	–	–	+	+	+	+	–	+	+	+	+	0	0	–	–	+	+	0	0	0	0
Tense	0	–	–	–	–	–	–	+	+	+	+	+	–	–	–	–	–	–	–	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nasal	0	–	–	–	–	+	+	0	0	0	0	0	–	–	–	–	–	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Continuous	0	0	0	0	0	0	0	–	+	+	+	+	0	0	0	0	0	0	0	0	0	0	–	+	+	–	–	–	+	0	0	0	0	0	0	0	0	0	0	0	0	0	–	+
Sonorant	0	–	+	–	+	0	0	0	–	+	–	+	–	+	–	+	+	0	–	+	–	+	0	–	+	–	–	–	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Palatalized	0	–	–	+	+	–	+	0	–	–	+	+	–	–	+	+	+	–	+	+	–	+	0	0	–	+	0	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0	–	+	–	+
Stress	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	–	+	–	–	+	–	–	+	–	–	+	0	0	0	0

§5. CLASSIFICATION ALGORITHMS (PART I)

Distinctive features and classifications

Classification algorithms permit the assessment of similarity and dissimilarity of linguistic phenomena. A linguistic unit is generally characterized by a certain selection of properties or distinctive features which are present in the particular phenomenon and are absent in others.

If these properties are given, the particular linguistic phenomenon can be described by assigning to it a vector of ones and zeros whose i-th coordinate corresponds to the i-th feature; it is equal to 1 if the particular object has the corresponding property and 0 otherwise.

It is sometimes assumed that the features may take on other values besides 1 and 0. In general, a distinctive feature is a certain numerical function defined on the set of the relevant objects.

If the feature may take on k values, it partitions the set of objects into at most k nonintersecting classes. Conversely, if there is a classification (partition) of the set of objects into k nonintersecting classes, one can introduce a feature which takes on k values. This inverse line of reasoning is characteristic of the decoding approach.

If two objects are described by the corresponding vectors, the similarity of the objects can be estimated by calculating the distance* between them as between points of n-dimensional space.

Table 4.2 /13/ characterizes the sounds of the Russian language.

The columns correspond to the phonetic letters of the Russian. A prime next to a consonant expresses soft pronunciation, a prime next to a vowel indicates that it is not stressed. The different features are listed in the horizontal rows. Vocality is assigned one of the two symbols + (vowel) or − (consonant) for each letter; the letter "j" in the author's opinion is neither a vowel nor a consonant, whereas "r," "r'," "l," "l'" are both vowels and consonants at the same time. Therefore, consonance is not specified by the value of vocality. Stress is a feature characteristic of vowels only, and for consonants it therefore takes on the value 0 (inapplicable).

The values + and − of a certain feature correspond to a greater similarity of sounds than + and 0 or − and 0, and the distance between the sounds may therefore be described by the function defined on p. 193.

If we want to apply the Euclidean distance $\rho_{euc} = \sqrt{(x_i - x'_i)^2}$, we should first assign a certain number to each value of the different features (0, +, and −). If we measure the distance between two sounds using equation (4.5) (p.), the distance between a and b will be 20, and the distance between a and o only 3. This agrees with the intuitive concept of similarity of sounds.

If we have a selection of so-called grammatical classes of words (e.g., "nominative case," "masculine," "singular"), we can construct an analogous table expressing the grammatical properties of words. Given a selection of classes of words with some common semantic denominator (e.g., animation, greatness, intelligence, etc.), we can construct a semantic description of words.

* Distance is a function of pairs of elements of a certain set with the following properties: 1) $\rho(a, a) = 0$ (nondegeneracy), 2) $\rho(a, b) = \rho(b, a)$ (symmetry), 3) $\rho(a, b) + \rho(b, c) \geq \rho(a, c)$ (the triangle inequality).

Besides providing a convenient means of assessing the similarity of objects, the description vectors can be used to replace the tremendous variety of objects with sequences comprising a limited number of distinctive features. Thus, using binary features, we can describe a set of n objects with the aid of $[\log_2 n]+1$ features.* This is a highly valuable property of the vector approach, seeing that the total number of various words and concepts is really enormous.

Examining Table 4.2, we note that the features cover a wide spectrum of properties: some of them are related to pronunciation, the others to acoustic properties of sounds.

If we were to construct a similar table from an analysis of the various combinations of sounds in fluent speech, we could reconstruct the sounds of the various letters from written text. After all, written language does not markedly distort the ability of sounds (as expressed by letters) to combine with one another. If similar tables were available for individual words, in such a way that classes of words corresponding to a certain value of each sign contained words with some common semantic denominator, we could "guess" the meaning of words from an examination of texts.

This problem encounters considerable difficulties. Therefore, the general scheme will help to better understand the classification algorithms described below.

Algorithms for the identification of vowels and consonants

The first algorithms of this class reconstruct the pronunciation of letters from the occurrence of their combinations in a text. It involves the partition of letters into two classes using a single binary feature.

If this algorithm is applied to letters, it will identify vowels and consonants; applying the algorithm to morphemes, we can distinguish between meaningful morphemes and auxiliary morphemes. Application of the algorithm to mathematical texts would differentiate between predicate symbols (+, −, =, etc.) and object symbols (e.g., π, x, 10, 2^{32}). When applied to words, the algorithm will probably differentiate between nouns and verbs.

By identifying the vowels and the consonants one naturally does not establish the exact pronunciation of the letters. However, this is a first useful step toward decoding.**

Thus, if the algorithm is applied to letters, it provides a definition of vowels and consonants. This definition is superior to conventional definitions (of acoustic or physiological bias) in that it is applicable to letters for which these traditional concepts are invalid.

The set of alternatives. The vowels and the consonants are thus regarded as a certain partition of the set of letters into two classes: the class of vowels and the class of consonants. In other words, it is assumed that these two sets are disjoint and between themselves exhaust the entire alphabet.

This restriction, however, does not quite correspond to the true state of things. Indeed, the letter "y" in English is sometimes rendered as a vowel and sometimes as a consonant (compare "very" and "year"). This is by no

* Here $[\log_2 n]$ stands for the whole part of the logarithm.

** The deciphering of inscriptions in the so-called Carian language carried out by V. V. Shevoroshkin began with the identification of vowels and consonants.

means a result of some imperfection in the written language (in Czech also, "r" is a consonant in the word "Praha" and a vowel in the word "prst," finger). If we do not restrict the analysis to letters, we see that this is a very common phenomenon; in particular, a single word often has a variety of meanings (the phenomenon of homonymy).

However, it would be impossible to develop an algorithm for the identification of vowels and consonants without these restrictions.

However, by stating that vowels and consonants constitute disjoint classes of a certain partition we have said very little. If a particular alphabet contains n letters, we may construct 2^n different partitions! Nevertheless, this statement is one step forward: so far the set of alternatives has not been restricted at all.

The quality function. The quality function is constructed from the following considerations: in any text, vowels are not very inclined to combine with other vowels and consonants with other consonants. Conversely, vowels readily combine with consonants.

If we take an arbitrary partition of the alphabet into two classes, we are not likely to notice this property. Suppose that the letter P has been declared as a consonant, and all the other letters of the alphabet as vowels. Under this partition, "vowels" may clearly occur very often in close combinations.

Let us analyze the combinations of letters of some language using a table whose rows and columns are identified by the letters of the corresponding alphabet. The entry corresponding to the row i and the column j contains a number which indicates how many times the letter a_i and the letter a_j occurred one next to the other in a given text (the order in which the two letters occurred is immaterial).

Consider a certain partition of the alphabet into two classes. All the rows and the columns headed by "vowels" are shifted to the left-hand top corner of the table, which is separated from the other letters by a line. The table thus takes the form

	vowels	consonants
vowels	1	2
consonants	4	3

Block 1 contains numbers which show how vowels combine with other vowels, block 3 contains numbers which show how consonants combine with consonants, and blocks 2 and 4 contain numbers showing how vowels combine with consonants. If the partition is close to the true division into vowels and consonants, the numbers in blocks 1 and 3 should be small, and those in blocks 2 and 4 large. The quality of the partition therefore can be estimated in terms of the sum of the numbers in blocks 1 and 3, say.

If the alphabet contains n letters, of which m are vowels, the corresponding quality function can be expressed in the form

$$K_1 = \sum_{i=1}^{m} \sum_{j=1}^{m} \varphi(a_i, a_j) + \sum_{k=m+1}^{n} \sum_{l=m+1}^{n} \varphi(a_k, a_l). \qquad (4.1)$$

Here $\varphi(a_i, a_j)$ is the number of joint occurrences of the letters a_i and a_j, regardless of order. The fact that we ignore the particular order in which the two letters combine is indicated by the comma, thus $\varphi(a_i, a_j) = \varphi(a_i a_j) + \varphi(a_j a_i)$. The letters a_i and a_j belong to one of the classes, and the letters a_k and a_j to another.

The smaller the value of K_1, the better is the partition. The best partition is that when the function is minimized.

The above estimate function for the detection of vowels is not the only possible one. Various equivalent estimate functions are available, which give an extremum for the same permissible solution which minimizes K_1. There are also interesting estimate functions which are not equivalent to K_1. One of these is

$$K_2 = \sum_{i=1}^{m} \sum_{j=1}^{m} p(a_i, a_j) p^2(a_x) + \\ + \sum_{k=m+1}^{n} \sum_{l=m+1}^{n} p(a_k, a_l) p^2(a_y) - \\ - 2 \sum_{i=1}^{m} \sum_{k=m+1}^{n} p(a_i, a_k) p(a_x) p(a_y). \qquad (4.2)$$

The symbol $p(a_x)$ stands for $\frac{\sum_{x=m+1}^{n} \varphi(a_x)}{N}$, where a_x belongs to the same class as a_k and a_l. This notation is based on the fact that the appearance of any letter of a given class can be regarded as the appearance of some letter a_x; similarly, $p(a_y)$ stands for $\frac{\sum_{y=1}^{m} \varphi(a_y)}{N}$. The number $p(a_y)$ is the relative frequency of one of the classes, and $p(a_x)$ the relative frequency of the other class.

The function K_2 is similar to the function K_3, which is equivalent to K_1,

$$K_3 = \sum_{i=1}^{m} \sum_{j=1}^{m} p(a_i, a_j) + \\ + \sum_{k=m+1}^{n} \sum_{l=m+1}^{n} p(a_k, a_l) - 2 \sum_{i=1}^{m} \sum_{k=m+1}^{n} p(a_i, a_k), \qquad (4.3)$$

differing from it in the coefficients $p(a_x)$ and $p(a_y)$. On the whole, the function K_2 is the correlation moment of the sequence of numbers 1 and -1 generated when 1 is substituted for each vowel in the text and -1 for each consonant. The function K_2 reflects the nonuniform frequencies of the vowels and the consonants. Combinations of consonants are more frequent, and therefore less significant; this is reflected in the weighting of the occurrenc' of consonants by the frequency of vowels, and vice versa. All this is highly hypothetical, however, since experiments were performed with the function K_1 only.

Recognition procedures. The simplest procedure based on the above features is quite trivial. It suffices to construct a table of the frequencies of pair combinations, examine all the possible partitions, and evaluate the estimate function for each partition. The partition corresponding to the minimum value of the quality function is then chosen. However, the volume of computations involved in this algorithm exceeds the ability of the largest modern computers for alphabets of normal size (e.g., about 30 letters).

The search for an effective procedure of minimizing the quality function is associated with considerable mathematical difficulties. The choice therefore lies between impracticable and incorrect methods.

We will describe an algorithm which often minimizes the function K_1 fast and without difficulty. For some tables it gives incorrect results, but even these apparently are not too far from the best solution. Anyway, experiments with this algorithm never led to errors which could be attributed to algorithm imperfection. This imperfection emerged only when specially selected tables were used.

We describe the procedure step by step:

1. For a given text, construct the table of the numbers $\varphi(a_i, a_j)$, where $\varphi(a_i, a_j)$ is the number of occurrences of the pair of letters a_i, a_j, irrespective of order.
2. Cross out numbers of the form $\varphi(a_i, a_i)$.*
3. Compute the sum of numbers in each row of the table.
4. Move to the first position (left-hand top corner) the row and the column with the largest sum.
5. Separate by vertical and horizontal lines the row and the column that were moved from the other rows and columns.
6. For rows below the horizontal boundary, calculate the sum of numbers lying to the right of the vertical boundary and the sum of numbers to the left of the vertical boundary; subtract the second sum from the first. The resulting numbers are called the decisive differences.
7. If there are positive decisive differences, move to 9.
8. End. Rows above the horizontal boundary correspond to letters of the first class (generally vowels), and those below the horizontal boundary correspond to letters of the second class.
9. Select the row with the maximum positive decisive difference and move it across the horizontal boundary; the corresponding column is moved across the vertical boundary. Return to 6.

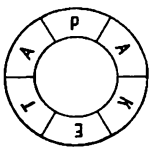


FIGURE 53. The word "ракета".

Consider a small example illustrating the application of this algorithm. Note that the algorithm can be used without separating successive words. In particular, suppose that we do not know which letter is the first letter of a word and which is the last letter (the word is inscribed along a circle, as in Figure 53).

The table of the numbers $\varphi(a_i, a_j)$ for this text has the form

	P	A	K	E	T
P		2			
A	2		1		1
K		1		1	
E			1		1
T		1		1	

* These numbers are the frequencies of pairs of identical letters. They clearly enter the sum $\sum_i \sum_j \varphi(a_i, a_j) + \sum_k \sum_l (a_k, a_l)$ for any partition of the alphabet into two classes, and therefore do not affect the quality of the particular partition.

After step 2, the table does not change. We proceed with step 3:

	P	A	K	E	T	
P		2				2
A	2		1		1	4
K		1		1		2
E			1		1	2
T		1		1		2

Now we proceed with steps 4 and 5:

	A	P	K	E	T
A		2	1		1
P	2				
K	1			1	
E			1		1
T	1			1	

Step 6:

	A	P	K	E	T	
A		2	1		1	
P	2					−2
K	1			1		0
E			1		1	2
T	1			1		0

From step 7 we move to step 9. Carrying out instructions 9 and 6, we get

	A	E	P	K	T
A			2	1	1
E				1	1
P	2				
K	1	1			
T	1	1			

From step 7 we move to 8 and end the analysis. The result shows that the first class contains the letters A and E, and the second class the letters P, K, T.

Table 4.3 illustrates the results of a similar machine experiment using Russian, English, and French texts of 10,000 words each.

The results for the Russian and the French texts are virtually error-free. Note that the Russian letters ь and ъ correspond to vowels which have long

TABLE 4.3. Russian

		Vowels											Consonants																						
		О	А	Е	И	У	Ъ	Ь	Я	Ы	Э	Ю	Б	В	Г	Д	Ж	З	Й	К	Л	М	Н	П	Р	С	Т	Ф	Х	Ц	Ч	Ш	Щ
Vowels	О	28	15	39	32	7	4	3	16	2	3	7	72	167	136	107	24	30	56	143	192	118	181	152	121	133	192	6	44	4	31	25	6	
	А	15	6	10	22	10	2	1	30	1	2	13	35	127	43	99	40	90	10	147	173	67	228	46	120	103	110	11	42	9	60	50	4	
	Е	39	10	28	57	7	5	15	5	19	4	4	46	83	37	93	44	35	34	43	166	87	202	71	156	129	124	0	16	19	63	22	27	
	И	32	22	57	52	13	9	2	36	4	2	2	23	95	25	60	26	37	31	83	148	102	143	42	68	89	95	5	28	18	43	39	21	
	У	7	10	7	13	0	3	2	4	1	3	19	20	20	26	42	12	11	0	41	24	26	43	27	42	34	53	0	10	2	16	16	8	
	Ъ	4	2	5	9	3	0	0	11	0	3	6	5	21	2	12	0	4	0	20	42	9	40	8	9	70	59	0	0	2	10	7	1	
	Ь	3	1	15	2	2	0	0	1	0	0	0	26	27	1	8	0	6	11	4	22	31	43	7	14	7	15	0	12	4	1	5	0	
	Я	16	30	5	36	4	11	1	8	0	0	0	8	31	6	15	4	6	1	10	30	10	46	14	11	60	31	1	4	0	1	1	1	
	Ы	2	1	19	4	1	0	0	0	0	0	0	14	13	1	11	0	4	10	3	13	15	37	8	8	5	14	0	6	3	0	4	0	
	Э	3	2	4	2	3	3	0	0	0	0	0	0	4	0	0	0	0	1	3	2	0	0	1	0	1	23	0	0	0	0	0	0	
	Ю	7	13	4	2	19	6	0	0	0	0	2	5	3	0	2	0	3	0	2	9	1	2	4	1	5	13	0	0	0	1	0	14	
Consonants	Б	72	35	46	23	20	5	26	8	14	0	5	0	2	0	1	0	1	2	7	12	0	3	0	9	5	2	0	4	0	0	0	2	
	В	167	127	83	95	20	21	27	31	13	4	3	2	8	5	24	1	20	5	17	16	16	16	17	21	44	21	0	8	4	3	21	0	
	Г	136	43	37	25	26	2	1	6	1	0	0	0	5	0	19	0	5	4	3	16	6	7	1	19	1	2	8	1	0	0	0	0	
	Д	107	99	93	60	42	12	8	15	11	0	2	1	24	19	2	13	10	2	4	9	5	39	4	24	7	10	0	0	8	0	0	0	
	Ж	24	40	44	26	12	0	0	4	0	0	0	0	1	0	13	0	2	0	12	4	0	17	0	10	1	0	0	0	0	0	0	0	
	З	30	90	35	37	11	4	6	6	4	0	3	1	20	5	10	2	0	3	6	10	14	20	3	5	2	6	0	2	1	0	0	0	
	Й	56	10	34	31	0	0	11	1	10	1	0	2	5	4	2	0	3	0	11	6	4	14	12	8	23	13	0	3	1	0	3	0	
	К	143	147	43	83	41	20	4	10	3	3	2	7	17	3	4	12	6	11	4	19	8	11	10	40	57	15	0	5	0	3	5	0	
	Л	192	173	166	148	24	42	22	30	13	2	9	12	16	16	9	4	10	6	19	2	10	13	21	4	43	8	0	9	0	5	12	0	
	М	118	67	87	102	26	9	31	10	15	0	1	0	16	6	5	0	14	4	8	10	16	23	10	12	21	5	0	2	1	3	0	0	
	Н	181	228	202	143	43	40	43	46	37	0	2	3	16	7	39	17	20	14	11	13	23	80	7	12	32	20	0	9	8	12	2	2	
	П	152	46	71	42	27	8	7	14	8	1	4	0	17	1	4	0	3	12	10	21	10	7	0	59	17	8	0	6	0	2	0	1	
	Р	121	120	156	68	42	9	14	11	8	0	1	9	21	19	24	10	5	8	40	4	12	12	59	0	16	48	0	20	1	2	4	0	
	С	133	103	129	89	34	70	7	60	5	1	5	55	44	1	7	1	2	23	57	43	21	32	17	16	16	160	0	6	0	12	1	0	
	Т	192	110	124	95	53	59	15	31	14	23	13	2	21	2	10	0	6	13	15	8	5	20	8	48	160	4	1	3	0	22	0	1	
	Ф	6	11	0	5	0	0	0	1	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
	Х	44	42	16	28	10	0	12	4	6	0	0	4	8	1	0	0	2	3	5	9	2	9	6	20	6	3	0	0	0	0	2	0	
	Ц	4	9	19	18	2	2	4	0	3	0	0	0	4	0	8	0	1	1	0	0	1	8	0	1	0	0	0	0	0	1	0	0	
	Ч	31	60	63	43	16	10	1	1	0	0	1	0	3	0	0	0	0	0	3	5	3	12	2	2	12	22	0	0	1	0	1	0	
	Ш	25	50	22	39	16	7	5	1	4	0	0	0	21	0	0	0	0	3	5	12	0	2	0	4	1	0	0	2	0	1	0	0	
	Щ	6	4	27	21	8	1	0	1	0	0	14	2	0	0	0	0	0	0	0	0	0	2	1	0	0	1	0	0	0	0	0	0	

TABLE 4.3 (cont.). French

		Vowels							Consonants																		
		e	a	o	i	u	y	k	b	c	d	f	g	h	j	l	m	n	p	q	r	s	t	v	w	x	z
Vowels	e	96	67	6	114	166	6	0	40	168	283	33	45	46	30	318	219	302	127	27	428	457	283	109	0	14	12
	a	67	16	0	116	90	22	1	27	63	81	34	37	25	8	173	72	166	127	5	194	144	118	89	0	7	1
	o	6	0	0	118	150	3	0	35	73	19	14	2	22	10	43	71	253	60	0	118	88	64	43	0	1	0
	i	114	116	118	4	68	1	4	13	45	43	17	12	8	5	137	55	100	17	5	85	147	140	39	0	10	1
	u	166	90	150	68	4	0	0	9	25	57	3	10	3	7	70	10	92	17	94	111	146	83	20	0	33	0
	y	6	22	3	1	0	0	0	0	4	1	0	0	0	0	11	11	3	7	0	0	7	2	0	0	0	0
	k	0	1	0	4	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3	0
Consonants	b	40	27	35	13	9	0	0	0	0	0	0	0	0	0	21	4	2	0	0	13	11	5	0	0	2	0
	c	168	63	73	45	25	4	1	0	16	2	0	0	56	0	15	1	55	3	0	27	37	36	0	0	3	2
	d	283	81	19	43	57	1	1	0	2	12	1	1	0	1	1	1	83	5	2	64	69	37	2	0	5	0
	f	33	34	14	17	3	0	0	0	0	1	10	0	0	0	11	0	11	0	0	7	5	3	0	0	1	0
	g	45	37	2	12	10	0	0	0	0	1	0	0	0	0	2	0	22	0	1	18	3	3	0	0	0	0
	h	46	25	22	8	3	0	0	0	56	0	0	0	0	0	8	0	4	2	0	1	8	1	0	0	2	0
	j	30	8	10	5	7	0	0	0	0	1	0	0	0	0	1	0	2	0	0	1	5	4	0	0	0	2
	l	318	173	43	137	70	11	0	21	15	1	11	2	8	1	74	3	24	28	7	36	88	44	3	0	2	2
	m	219	72	71	55	10	11	0	4	1	1	0	0	0	0	3	62	5	19	0	12	9	11	0	0	1	0
	n	302	166	253	100	92	3	0	2	55	83	11	22	4	2	24	5	60	14	14	16	100	170	12	0	0	0
	p	127	127	60	17	17	7	0	0	3	5	0	0	2	0	28	19	14	22	0	58	49	28	0	0	7	2
	q	27	5	0	5	94	0	0	0	0	2	0	1	0	0	7	0	14	0	0	8	20	6	0	0	1	0
	r	428	194	118	85	111	0	1	13	27	64	7	18	1	1	36	12	16	58	8	96	58	160	10	0	0	0
	s	457	144	88	147	146	7	1	11	37	69	5	3	8	5	88	9	100	49	20	58	152	112	11	0	1	0
	t	283	118	64	140	83	2	0	5	36	37	3	3	1	4	44	11	170	28	6	160	112	54	10	0	1	0
	v	109	89	43	39	20	0	0	0	0	2	0	0	0	0	3	0	12	0	0	10	11	10	0	0	2	2
	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	x	14	7	1	10	33	0	3	2	3	5	1	0	2	0	2	1	0	7	1	0	1	1	2	0	0	0
	z	12	1	0	1	0	0	0	0	2	0	0	0	0	2	2	0	0	2	0	0	0	0	2	0	0	0

TABLE 4.3 (cont.). English

		Vowels							Consonants																		
		e	o	a	i	t	u	y	b	c	d	f	g	h	j	k	l	m	n	p	q	r	s	v	w	x	z
Vowels	e	116	48	112	63	108	8	52	70	65	204	57	56	430	1	35	135	109	154	61	1	312	170	111	106	7	1
	o	48	114	15	20	157	115	17	29	39	71	97	41	91	0	38	119	70	125	30	0	139	80	14	91	0	0
	a	112	15	6	48	138	4	44	24	40	89	27	64	97	0	16	104	87	214	34	1	105	170	30	90	1	1
	i	63	20	48	0	102	9	8	5	26	73	37	39	140	0	23	88	48	203	64	8	119	121	23	53	0	1
	t	108	157	138	102	104	57	32	8	13	60	35	20	352	1	7	43	16	119	19	0	77	139	1	28	0	0
	u	8	115	4	9	57	0	0	32	9	13	7	34	12	3	0	28	14	63	14	12	31	52	0	0	0	0
	y	52	17	44	8	32	0	2	18	5	17	8	5	20	0	2	42	16	13	5	1	33	25	3	19	0	1
Consonants	b	70	29	24	5	8	32	18	4	0	15	1	3	6	1	1	7	6	11	1	0	18	13	0	1	0	0
	c	65	39	40	26	13	9	5	0	0	5	1	1	46	0	11	8	1	16	1	0	14	23	0	1	1	0
	d	204	71	89	73	60	13	17	15	5	22	11	12	49	2	0	63	20	155	5	1	44	39	1	17	0	0
	f	57	97	27	37	35	7	8	1	1	11	16	4	12	0	2	27	2	5	4	0	18	20	0	7	0	0
	g	56	41	64	39	20	34	5	3	1	12	4	6	45	0	1	2	4	106	1	1	31	25	0	9	0	0
	h	430	91	97	140	352	12	20	6	46	49	12	45	16	0	3	4	5	27	8	0	23	62	0	66	0	0
	j	1	0	0	0	1	3	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	k	35	38	16	23	7	0	2	1	11	0	2	1	3	0	0	13	0	26	1	0	7	15	0	1	0	0
	l	135	119	104	88	43	28	42	7	8	63	27	2	4	0	13	138	8	16	33	0	7	30	3	15	0	0
	m	109	70	87	48	16	14	16	6	1	20	2	4	5	0	0	8	2	5	6	0	11	17	1	8	0	0
	n	154	125	214	203	119	63	13	11	16	155	5	106	27	0	26	16	5	20	0	0	17	44	1	32	0	0
	p	61	30	34	64	19	14	5	1	1	5	4	1	8	0	1	33	6	0	26	0	20	23	13	10	5	0
	q	1	0	1	8	0	12	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0
	r	312	139	105	119	77	31	33	18	14	44	18	31	23	0	7	7	11	17	20	1	46	45	2	10	0	0
	s	170	80	170	121	139	52	25	13	23	39	20	25	62	0	15	30	17	44	23	1	45	74	7	41	0	0
	v	111	14	30	23	1	0	3	0	0	1	0	0	0	0	0	3	1	1	13	0	2	7	0	0	0	0
	w	106	91	90	53	28	0	19	1	1	17	7	9	66	0	1	15	8	32	10	1	10	41	0	6	0	0
	x	7	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
	z	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

since lost their vocality in the living language. The French letter k occurs very seldom, mainly in abbreviations (e.g., in initials of non-French names).

The error in the English-language table is associated with the use of the letter "t" in combinations which represent a distinct phonetic sound. An algorithm correcting errors of this kind and leading to successful results is described in /12/.

One of the "mathematically correct" algorithms minimizing K_1 is given on p. 187. A related algorithm which converts the so-called "syllabic writing" into normal letter writing is given on p. 188, and an algorithm identifying classes of words with a common meaning is described on p. 192.

§6. MATCHING ALGORITHMS (PART I)

Algorithms identifying code sequences

Algorithms intended for the detection of larger textual units, when the smaller elements are known, evidently constitute one of the most important classes of recognition algorithms.

We start our description of these algorithms with one of the simplest: an algorithm identifying letter codes by uniform-length sequences of symbols.

The importance of this problem for the case of extraterrestrial communication is obvious. The "elementary signal" of a message transmitted by an extraterrestrial civilization may have a simple form, in particular representable as one of the two binary symbols, 0 and 1. To transmit a longer alphabet, coding will have to be used, representing letters by sequences of the elementary signals. These signals quite likely may be of uniform length for all the letters of the alphabet.

The set of permissible solutions (the set of interpretations) in this case is found without difficulty. Let m be the length of the code groups, and N the length of the text expressed in elementary symbols. The number of permissible solutions in this case is m: it is determined by the number of shifts of the text through i digits ($i = 0, 1, \ldots, m-1$). Cyclic arrangement is assumed, whereby the last letter of the text is followed by the first letter.

If N and m are relatively prime numbers, the residue obtained in the division of N by m will be omitted. Therefore, the total number of permissible solutions is $\sum_{m=1}^{N/2} \left[\frac{N}{m}\right]$, where $\left[\frac{N}{m}\right]$ is the whole part of the corresponding quotient. This number is not greater than

$$\sum_{m=1}^{N/2} \frac{N}{m} < \frac{N^2}{2} \int_1^{N/2} \frac{dm}{m} = \frac{N^2}{2} \ln \frac{N}{2}.$$

Let us now proceed with a discussion of the quality function. Consider a text of length N encoded by groups of numbers of length m. What distinguishes this text from a random number sequence partitioned into blocks of the same length m? It is obvious that the frequencies of the m-letter groups in the second case should be much more uniform than in the first case. After

all, the second number sequence has been picked up "at random," and none of the numbers has any preference over other numbers.

On the other hand, the selection of letters in an ordinary message is far from random. There are sounds and sequences of sounds which are relatively easy to pronounce; if the message alphabet is the set of words in the message, different words occur with different frequencies, because of considerations of "common usage" and depending on the meaning of the message.

If the encoded text contains groups of length m, and we attempt to interpret them as containing groups of length $p (p \neq m)$ or at least groups of length m but displaced through i positions (where i and m are relatively prime), the message becomes similar to the sequence of symbols obtained by "repeating selection." This means that the elements of the code group corresponding to a single letter are more intimately related than the elements which belong to different code groups.

This sounds reasonable because incorrect "partitioning into groups" is devoid of those "preference criteria" which restricted the letter combinations.

It is therefore natural to use a quality function which reaches an extremum for a uniform distribution of the code element frequencies and also for a certain "highly" nonuniform frequency distribution.

Unfortunately, intuitive reasoning is not enough for an a priori choice of a quality function assessing diversity.

A whole range of traditional evaluation techniques are known. These include, for instance, the calculation of the root mean square deviation, the modulus variance, the entropy.

Our calculations based on a limited text pointed in favor of the function

$$V = \sum_i (\varphi(c_i) - \overline{\varphi(c)})^2.$$

Here c_i is a certain group of a given length, $\overline{\varphi(c)}$ is the mean absolute frequency of a group of this length, equal to $\frac{N}{m \cdot |A|^m}$, where m is the chain length, N is the length of the text in unit symbols, $|A|$ is the number of letters in the alphabet of unit symbols, $|A|^m$ is the number of letters in the alphabet of groups of length m, $\frac{N}{m}$ is the number of groups of length m in the given text (rounded off); thus V is the sum of the squares of the deviations of the actual frequencies from the mean absolute frequencies of the groups.

For a given group length, the function V is minimum when all the group frequencies are equal (then $V = 0$) and maximum when one symbol recurs through the entire text.

To permit comparison of the results for various group lengths, the expression

$$\sum_i (\varphi(c_i) - \overline{\varphi(c)})^2$$

is multiplied by a normalizing factor ν. This factor can be calculated if we proceed from the assumption that the best (maximum) value of V should be independent of m. Since the best (from the point of view of the particular

function) solution involves a single element only, its frequency is $\frac{N}{m}$, and the frequencies of the other elements are zero. Then V is equal to

$$\left(\frac{N}{m}-\frac{N}{m\cdot|A|^m}\right)^2+\left(\frac{N}{m\cdot|A|^m}\right)^2(|A|^m-1)=$$
$$=\frac{N^2}{m^2}\left[\left(1-\frac{1}{|A|^m}\right)^2+\frac{|A|^m-1}{|A|^{m\cdot 2}}\right]=\frac{N^2(|A|^m-1)\cdot|A^m|}{m^2|A|^{m\cdot 2}},$$

and since usually $|A|^m$ is large, we may take $|A|^m-1\approx|A|^m$, so that $V_{max}\approx\frac{N^2}{m^2}$.

Let the group length in some other solution be l; the maximum value of V is then approximately equal to $\frac{N^2}{l^2}$.

The normalizing factor is introduced so that the best values are equal:

$$\frac{N^2}{m^2}=\nu\frac{N^2}{l^2}.$$

Hence

$$\nu=\frac{l^2}{m^2}.$$

In the example that follows, a partition into groups of length $m=3$ is used as the "basis for comparison"; for 3-digit groups, the coefficient ν is equal to 1.

A short English text* has been encoded by a sequence of three-digit numbers using the following table:

$a=000$	$j=100$	$s=200$
$b=001$	$k=101$	$t=201$
$c=002$	$l=102$	$u=202$
$d=010$	$m=110$	$v=210$
$e=011$	$n=111$	$w=211$
$f=012$	$o=112$	$x=212$
$g=020$	$p=120$	$y=220$
$h=021$	$q=121$	$z=221$
$i=022$	$r=122$	

The last three-digit group is not used. The encoded text will look as follows:

201 021 011 201 211 022 020 021 102 022 020 021
201 022 200 200 000 010 000 111 010 002 102
112 202 010 220 201 021 011 211 022 111 010
001 102 112 211 200 211 022 102 010 000 111 010
012 122 011 011 000 111 010 000 200 201 021 011
211 022 111 020 200 112 012 200 011 000 001 022
122 010 200 012 102 000 200 021 201 021 011 211
021 022 201 011 002 000 120 200 112 012 201 021 011 200 011 000

* The first stanza of R. L. Stevenson's poem "Twilight":

The twilight is sad and cloudy,
The wind blows wild and free,
And as the wings of sea birds
Flash the white caps of the sea.

Table 4.4 lists the absolute frequencies of the three-digit groups for partitions beginning with the first, second, and third letter of the text, respectively (R_1^3, R_2^3, and R_3^3).

TABLE 4.4

R_1^3	200	011	000	021	010	022	201	211	102	111	012	112	020	001
	10	10	9	9	8	8	8	6	5	5	4	4	3	2
R_2^3	110	010	000	001	002	100	112	210	020	122	220	221	021	202
	15	8	7	6	6	6	6	6	5	4	4	4	3	3
R_3^3	102	101	020	000	201	001	011	120	220	100	121	211	002	202
	14	11	9	7	6	5	5	5	5	4	4	4	3	3
R_1^3	002	122	120	202	220	100	101	110	121	210	212	221	222	
	2	2	1	1	1	0	0	0	0	0	0	0	0	
R_2^3	102	120	121	200	212	222	011	012	211	022	101	111	201	
	2	2	2	2	2	2	1	1	1	0	0	0	0	
R_3^3	210	110	200	212	012	021	022	221	010	111	112	122	222	
	3	2	2	2	1	1	1	1	0	0	0	0	0	

The factor ν is taken equal to unity. We find

$$V(R_1^3) = 340.30; \quad V(R_2^3) = 277.34, \quad V(R_3^3) = 321.37.$$

The absolute frequencies for two-digit groups and partitions beginning with the first (R_1^2) and the second (R_2^2) letter of the text are given in Table 4.5.

TABLE 4.5

R_1^2	10	00	20	01	11	02	21	22	12
	26	24	19	17	15	14	12	11	9
R_2^2	02	00	01	11	10	20	21	12	22
	25	23	21	17	16	16	12	11	6

Taking for the normalizing factor $\nu = \frac{2^2}{3^2} = 0.44$, we find $V(R_1^2) = 268.00 \cdot 0.44 = 117.92$, $V(R_2^2) = 295.44 \cdot 0.44 = 129.99$. Both figures are markedly less than $V(R_1^3)$. The absolute frequencies for $m = 1$ are

$$\varphi(0) = 124; \quad \varphi(1) = 94; \quad \varphi(2) = 76.$$

The normalizing factor is equal to $\frac{1}{9}$; we thus have $V(R_1^1) = 1176 \cdot \frac{1}{9} = 130.67$. This is again less than $V(R_1^3)$.

Consider the absolute frequencies for four-digit groups. In this case, the partitions may start with the first, second, third, and fourth letter of the text (R_1^4, R_2^4, R_3^4, R_4^4, respectively).

The frequencies of the symbols are listed in Table 4.6.

TABLE 4.6

R_1^4

Absolute frequency	Number of groups
5	1
4	2
3	4
2	12
1	24
0	38

R_2^4

Absolute frequency	Number of groups
5	2
3	5
2	12
1	24
0	38

R_3^4

Absolute frequency	Number of groups
4	1
3	9
2	9
1	24
0	38

R_4^4

Absolute frequency	Number of groups
4	2
3	7
2	13
1	27
0	35

The normalizing factor is $\frac{4^2}{3^2} = \frac{16}{9} = 1.67$, and $V(R_1^4) = 165.77$, $V(R_2^4) = 169.02$, $V(R_3^4) = 141.38$, $V(R_4^4) = 125.88$.

To establish that V is indeed maximum for the correct partition, we should calculate the values of this function for all $m\left(m = 1, 2, \ldots, \frac{N}{2}\right)$. This is, however, not absolutely essential: clearly, the squares of the differences $(\varphi(c_i) - \overline{\varphi(c)})^2$ markedly decrease as m increases, whereas ν increases only moderately.

Another thought is that groups longer than twenty elements need not be considered altogether; after all, even assuming a binary set of elementary symbols, the power of the alphabet of groups of this length is 2^{20}, i.e., more than enough to represent the most complex alphabets (including the Chinese).

Let us compare the entropies calculated for some of these partitions.

The entropy $H = -\sum_i p_i \log p_i$ for a uniform distribution is maximum; it is zero if the probability of one element is 1 and of all the others 0. We replaced the probability p with the modified quantity $p = \frac{\varphi(c_i)}{\frac{N}{m}}$, where $\frac{N}{m}$ is the length of text in terms of m-digit groups.

We have $H(R^3) = 1.165$, $H(R_2^3) = 1.248$, $H(R_3^3) = 1.226$, which is again quite satisfactory. For all other lengths of code groups, the entropy should be normalized by dividing by $\log m$. It is remarkable that the V corresponding to almost all incorrect partitions have close values: this again proves the adequacy of normalization.

An example illustrating the application of the concept of meaningfulness

The above algorithm can be applied to demonstrate the concept of meaningfulness, presented in §2. Let us identify the set of interpretations of the

text with the set of permissible solutions of the given logarithm, defining meaningfulness as the difference of the functions V or H for the worst and the best partition.

The results which can be obtained following this approach are best illustrated by an example. Consider a binary alphabet of elementary signals, with two elements 0 and 1. The code group is 2 elements long. In this case, the set of interpretations contains only two partitions, R_1^2 with the groups starting with the odd elements, and R_2^2 with the groups starting with the even elements.

How do these interpretations look in the best case, when $\Delta(H) = H(R_1^2) - H(R_2^2)$ is maximum? Let us characterize the partition by the probability distributions of the code groups:

	Partition	
	R_1^2	R_2^2
Distributions	$p'(00)$	$p''(00)$
	$p'(01)$	$p''(01)$
	$p'(10)$	$p''(10)$
	$p'(11)$	$p''(11)$

These probabilities are not independent. They can be expressed in terms of the probabilities of 4-digit groups beginning, say, with elements whose running number in the text is a multiple of 4:

$$
\begin{aligned}
p'(00) &= p(0000) + p(0001) + p(0010) + p(0011)\\
p'(01) &= p(0100) + p(0101) + p(0110) + p(0111)\\
p'(10) &= p(1000) + p(1001) + p(1010) + p(1011)\\
p'(11) &= p(1100) + p(1101) + p(1110) + p(1111)\\
p''(00) &= p(0000) + p(0100) + p(1000) + p(1100)\\
p''(01) &= p(0001) + p(0101) + p(1001) + p(1101)\\
p''(10) &= p(0010) + p(0110) + p(1010) + p(1110)\\
p''(11) &= p(0011) + p(0111) + p(1011) + p(1111)
\end{aligned}
$$

Inserting in the expression for $\Delta(H)$ the values of p' and p'' expressed in terms of $p(\alpha\alpha\alpha\alpha)$, we obtain a function of 16 variables, whose maximum will enable us to compute the two distributions. The distributions $\{p_i'\}$ and $\{p_i''\}$ characterize a text of maximum meaningfulness in a certain sense! Taking averages of the form $p_i^{(2)} = \frac{p_i' + p_i''}{2}$, we obtain the probability distribution of the pairs of elementary symbols of the text, and taking the sums $p(01) + p(00)$ and $p(01) + p(11)$ we obtain the probability distribution of the one-digit symbols $\{p_i^{(1)}\}$. These distributions permit computation of the second-order entropy of the text using the equality $H_2 = H(p_i^{(2)}) - H(p_i^{(1)})$, and hence the approximate redundancy (or, more precisely, the lower limit of redundancy) as $1 - \frac{H_\infty}{H_0}$, replacing H_∞ with H_2 and H_0 with $\log 2$.

Highly interesting results could be obtained if the "intelligibility" were evaluated in terms of entropy of infinite order. Calculations carried out under this assumption would help to estimate the level of entropy and redundancy characteristic of meaningful texts.

§7. PATTERN DECODING ALGORITHMS

The language of images. Connectedness and detailedness

Our definition of intelligibility is neither widespread nor usual. Usually, a text is considered intelligible if it produces mental association with some real situations or images.

This approach naturally does not answer the question why certain situations from reality are unintelligible. Nevertheless, the usual interpretation of intelligibility is largely valid. After all, there is a fundamental correlation between the "predictive systems" of the human language and the human reality, or, to use a different turn of phrase, words combine roughly in the same manner as the real phenomena that they represent.

Thus if man speaks, moves, and interacts with the surrounding objects, the word "man" will naturally also combine with words designating speech, motion, action. Another remarkable correspondence is observed between sentences, which are generally made up of words designating objects (nouns), actions (verbs), and properties (adjectives), and typical real situations which are made up of objects, their interactions, and properties.

This correspondence is far from trivial, and yet it is not too complicated, so that no special translation rules had to be devised in any of the languages for one particular situation.

The translation from the language of reality to human language and back is naturally a very complicated undertaking; there is, however, one peculiar human language for which this translation is done without much difficulty. We mean here the language of images.

This language clearly suffers from considerable shortcomings. It is highly uneconomic: e.g., compare the sentence "a man walks" with a picture announcing the same fact. A correct image must contain a great wealth of detail, which is often immaterial for the case being considered. Moreover, some messages do not lend themselves to translation into the language of images without sacrificing the simplicity of the mapping which relates the image to the real situation (e.g., such sentences as "perseverance wins" or "1963 was a droughty year").

There is therefore no reason to suggest that the language of images would be the only means of interstellar communication.

However, its great advantage is its intelligibility. It is not only that the image language can be readily translated into the usual language of reality: there is a very strong predictive relationship between the adjacent elements of an image.

The graphic form of the decoding problems associated with image analysis is another highly favorable feature, enabling us to consider these problems as models for the more difficult task of decoding of the ordinary language.

One of the typical decoding problems in the analysis of image languages is the following: consider a sequence of signals; it is required to convert it into a two-dimensional picture so that an intelligible image message is obtained. A typical feature of this problem is that it gives rise to serious doubts concerning the usefulness of a formal definition of "meaningfulness." After all, the human mind will immediately distinguish between a meaningful and a meaningless picture.

Let the sequence of signals comprise the lines of a rectangular scan of an image consisting of black and white dots (represented by the digits 1 and 0, respectively) arranged in succession. This sequence can be decoded in the following way: by changing some parameter d from 1 to N (N is the sequence length), we partition the sequence into lines of length d and arrange these lines one under the other.

A man examining the picture formed in the process will instantly recognize the best of the various images.

The reader may wish to experiment on his own with the following cosmogram:

```
    1001111110010011111110000000001
                                    00
1111111000001110000000111110000 0
1
1001011111010010111110100101111 1
                                   0
1111101001011111010010111110100 1
0
1000011011000000011011000000011011
                                   0
1101100000011011000000011011000000
0
0010011011001
```

The decoding of this message is

```
1001111111001
0011111111100
0001010101000
0001111110000
0000111100000
0111111111110
0101111110101
0101111110101
0101111110101
0101111110101
0101111110101
0101111110101
0001101101000
0001101101000
0001101101000
0001101101000
0001101101000
0001101101000
1001101101001
```

and the ones form the picture of a man in a hat.

This method of image decoding on a rectangular screen will possibly be the most effective if a computer is entrusted with the task of deriving the set of all possible solutions (i.e., the partition of the text into segments of length l and their arrangement one under the other). The situation radically

changes, however, if we consider messages obtained by scanning a screen of an arbitrary shape. In this case, the number of permissible solutions increases prohibitively.

Even a computer will not be able to examine and assess all the permissible solutions in this case. However, if the computer has been programmed with a formal criterion of meaningfulness, it may apply a shorter procedure proceeding, say, from a somewhat less meaningful image or part of an image to a more meaningful one. The difference in meaningfulness between the two successive images may be so slight as to be actually imperceptible to the human eye.

Consider two image scanning techniques: the image is covered by two systems of x and y coordinate lines, and the lines of each system are spaced a certain distance $\Delta\rho$. The part of the image between two adjacent lines of the system x will be called a line, and the part of the image between two successive lines of the system y will be called a column. The element of the text located between adjacent lines of the systems x and y will be called a dot. If a certain classification of the dots is given, each dot may be replaced by the corresponding classification digit; the lines of digits are then numbered and arranged in sequence one after the other.

A preliminary hypothesis for the construction of an image quality criterion presupposes that the system of coordinate lines always can be defined in an optimum manner for the particular image.

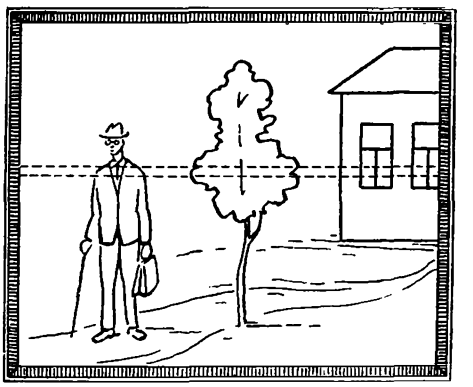


FIGURE 54. In this picture, rich in vertical details, the "horizontal lines" are highly similar.

In this, we have to lean on a certain property of texts, which is apparently fairly general. Meaningful messages probably fall into component parts, not unlike sentences, which in their turn are composed of smaller elements (analogous to words). The "quasiwords" in the "quasisentences" should be different in a certain sense; the adjacent "quasisentences," on the other hand, should be close to each other in a sense.

Thus, Russian-language sentences are made up of words expressing a variety of concepts: nouns signifying objects, verbs signifying action, processes, or states, adjectives qualifying properties. The real images or

situations corresponding to the different words in a sentence are also highly different. At the same time, nearby sentences have similar structure and often close meaning, which imparts the sense of "connectedness" to the text. For instance, the sentences "A vessel emerged from beyond the horizon. This was a boat with a wide white stack" deal with a common subject, expressed by the words "vessel" and "boat," and they thus appear "connected."

These properties are even more prominent in a picture: if Figure 54 is cut into horizontal lines, the black and the white dots will frequently alternate; adjacent lines will moreover be very similar to each other, whereas in columns the black and white dots alternate infrequently.

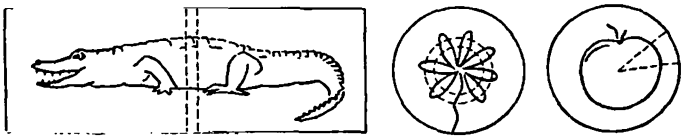


FIGURE 55. Pictures rich in horizontal, radial, and concentric lines.

This is clearly true only if the partition into lines and columns is done according to a certain pattern: thus, for the picture of a crocodile the lines should be vertical and the columns horizontal; for the picture of a flower, the lines are concentric circles, and for the picture of an apple they are radial lines (Figure 55).

A similar property is characteristic of messages composed in formal languages and in LINCOS-type languages: adjacent sentences in these languages are "logically sequential" and they are generally similar to one another when presented in graphical form.

For fairly complex images, the choice of correct coordinate lines is apparently not so significant, because for any direction of the lines, two adjacent lines will have a similar appearance and will contain a frequent alternation of black and white dots. We will say that similarity of adjacent scanning lines ensures connectedness of images, similarity of more distant lines ensures smoothness, and variety within the lines ensures detailedness.

Examples of quality functions. Some procedures

A simple quality function can be proposed evaluating images in terms of connectedness and detailedness. Detailedness is assessed as the number of transitions from a black dot to a white dot within a single line, and connectedness as the number of black—white transitions occurring in corresponding positions in two adjacent lines. Let 1 stand for a black dot and 0 for a white dot. The function $u_{i,i+1}$ assessing the quality of adjacent lines can be written in the form

$$u_{i,i+1} = \varphi(|01) + \varphi(|10),$$

where $\varphi(|01)$ is the number of transitions from a white to a black dot in the $(i+1)$-th line occurring below identical transitions in the i-th line, $\varphi(|10)$ is

the number of such binary transitions from a black to a white dot. The line-wise image quality is expressed in the form

$$U_{\text{line}} = \sum_i u_{i,\,i+1}.$$

Since we do not know in advance if the "quasisentences" are lines or columns, column-wise quality function should also be evaluated, identifying closeness of adjacent columns according to the formula

$$u'_{j,\,j+1} = \varphi(\overline{01}) + \varphi(\overline{10}),$$

where $\varphi(\overline{01})$ is the number of white-to-black transitions along the vertical, situated next to the corresponding transitions in the column immediately to the left. The column-wise image quality is then expressed by the equality

$$U_{\text{column}} = \sum_j u'_{j,\,j+1},$$

and the overall image quality U is given by

$$U = \sum_i u_{i,\,i+1} + \sum_j u'_{j,\,j+1}. \tag{4.4}$$

The summation can be carried out along the lines only, since the sum $\sum_j u'_{j,\,j+1}$ is equal to the sum $\sum u'_{i,\,i+1}$, where $u'_{i,\,i+1} = \varphi(\overline{01}) + \varphi(\overline{10})$ in adjacent lines.

It is not entirely clear how to treat the first and the last symbol in adjacent lines. In our conception, the partition into lines is equivalent to introduction of special "boundary symbols." The transition to a boundary symbol naturally carries certain information and should affect the image quality if it is transmitted together with the image. However, since by assumption the image being decoded does not contain special boundary symbols, this information is "fictitious" and should be minimized.

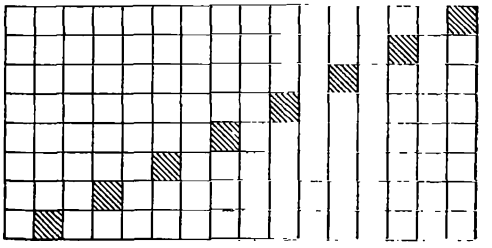


FIGURE 56. The picture of an inclined solid line obtained with a square grid.

For simplicity, we will assume that the image is unbounded in both the vertical and the horizontal direction, i.e., it is drawn on a torus (a steering-

wheel); the last line is followed by the first line, and the right-most column is followed by the left-most column. If the last line is partially filled, it is completed with a more frequent element, e.g., with zeros.

Consider some of the first decodings of the text

0000001010010100111000010000 1000000

with the respective quality functions (for the other decodings, only the quality function is given):

Length of line	1	2	3	4	5	6	7
Interpretation	0	00	000	0000	00000	000000	0000001
	0	00	000	0010	01010	101001	0100101
	0	00	101	1001	01010	010011	0011100
	0	10	001	0100	01110	100001	0010000
	0	10	010	1110	00010	0000010	1000000
	0	01	011	0001	00010	000000	
	1	01	100	0000	00000	$U=2$	$U=6$
	0	00	001	1000	$U=10$		
	1	11	000	0000			
	0	10	100	$U=4$			
	0	00	000				
	1	01	000				
	0	00	$U=4$				
	1	00					
	0	10					
	0	00					
	1	00					
	1	00					
	1	$U=6$					
	0						
	0						
	0						
	0						
	1						
	0						
	0						
	0						
	0						
	1						
	0						
	0						
	0						
	0						
	0						
	0						
	$U=0$						

Length of line	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
U	6	4	2	6	0	6	4	4	8	4	8	2	4	8	4	8	4	4	4

Length of line	26	27	28	29	30	31	32	33	34	35
U	4	4	4	4	4	4	4	4	4	0

A correct interpretation is the decoding with a five-element line (the picture of the numeral 4). This line length also corresponds to the maximum value of the quality function $U = 10$.

The above quality function is suitable for images rich in thin and solid vertical or horizontal lines. However, it will give erroneous results for discontinuous images, and for images with prevalent diagonal lines. Both cases are interrelated; indeed, a square grid cannot form a continuous image of a diagonal line (while preserving the line width). Let us examine Figure 56.

In this figure, all the centers of the black squares lie along the straight line $y = \frac{1}{2}x + \frac{1}{4}$. It is readily seen that not a single additional square can be hatched without breaking this condition. The quality function avoiding this difficulty makes use of what is known as image smoothness. We define a special operation, called "linear forecasting," which assigns a third line λ_p to any pair of lines λ_i and λ_j.

This operation is carried out as follows:

The elements of the line λ_i are joined with the elements of the line λ_j by straight segments observing the following three conditions:

1. Every element of the line λ_i is joined at least with one element of the line λ_j.
2. Every element of the line λ_j is joined at least with one element of the line λ_i.
3. When conditions 1 and 2 are observed, the sum of the segment lengths is minimum.

The segments are then continued to an arbitrary distance. If the segments are continued to row $j+k=t$, we say that the maximum forecast depth is k. A forecast of depth k is implemented as follows: the squares with the segments passing through their centers are identified as black squares (ones) and all the other squares remain white (zeros). For a rectangular screen, the length of the line q_p is the same as the length of the lines q_i and q_j (in our example, $k=1$) (Figure 57).

In other than rectangular screens, the position of the boundary points is first determined (Figure 58).

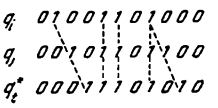


FIGURE 57. Forecast of depth 1 on a rectangular screen.

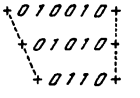


FIGURE 58. Forecast of depth 1 on a screen of arbitrary shape.

The resulting line is then compared with a real line λ_t occupying the same position, using the function $u_{r,t}$.

In practice, it is probably always sufficient to compare two adjacent lines and to make a forecast of depth 1, i.e., to forecast the adjacent line.

If the number $u_{r,t}$ obtained from a forecast using the lines λ_i and λ_{i+1} is designated $u^*_{r,t}(\lambda_{t,\,i+1})$, the image quality may be estimated with the function

$$U^*_{\text{line}} = \sum_{i=1}^{n} u^*_{r,t}(l_{t,\,i+1}).$$

Example. For the pattern

0 0 0 0 0 0 0
0 1 0 0 0 1 0
0 0 1 0 1 0 0
0 0 0 1 0 0 0
0 0 1 0 1 0 0
0 1 0 0 0 1 0
0 0 0 0 0 0 0

$U^{*}_{\text{line}} = 10$, which is much better than $U_{\text{line}} = 0$.

Similarly to linear forecasting, we could define nonlinear forecasting, which uses three lines to reconstruct a fourth. For example,

0 1 0 0 0 0 1 0 0 0
0 1 0 0 0 0 1 0 0 0
0 0 1 0 0 0 0 1 0 0
0 0 0 0 1 0 0 0 0 1

Here the difference between two adjacent horizontal shifts is preserved.

These techniques of bypassing the difficulties associated with discontinuity are logically irreproachable and do not look excessively arbitrary. A simpler method is described in the following.

The presence of diagonal lines with squares touching at the corners may be allowed for if the lines in (5.4) are replaced with bottom-to-top diagonals and the columns with top-to-bottom diagonals. Designating a pair of transitions along adjacent diagonal "lines" as /01 or /10, and a pair of analogous transitions in the diagonal "columns" as \01 or \10, we may define the similarity of diagonal lines and columns as $u^{\text{diag}}_{l,\,l+1} = \varphi(/10) + \varphi(/10)$ and $u'^{\text{diag}}_{l,\,l+1} = \varphi(\backslash 01) + \varphi(\backslash 10)$ and the "diagonal quality function" as $U^{\text{diag}} = \sum_{l} u^{\text{diag}}_{l,\,l+1} + \sum u'^{\text{diag}}_{l,\,l+1}$.

The total image quality is expressed as the sum of the two "quality" functions:

$$U^{\text{tot}} = U + U^{\text{diag}}.$$

For fairly complex images, however, the function u is quite sufficient.

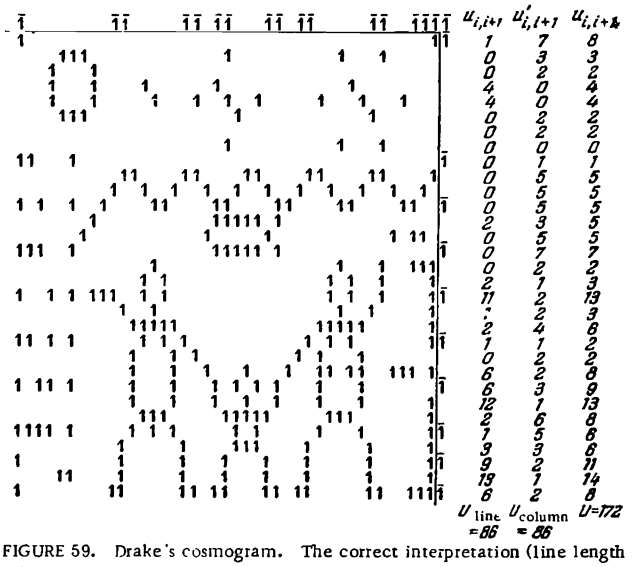


FIGURE 59. Drake's cosmogram. The correct interpretation (line length 41). The first line is preceded by the last line and the last column is followed by the first column.

Figure 59 and Table 4.7 show Drake's cosmogram. The values of $u_{i,\,i+1}$; $u'_{i,\,i+1}$ and the corresponding sums are also given. Note the considerable difference in the values of u for the correct and incorrect interpretation. Both pictures are assumed to be drawn on a torus.

TABLE 4.7. Drake's cosmogram. Incorrect interpretation (line length 64). The quality (142) is much lower than the quality of the correct interpretation (172).

		$u_{i,\,i+1}$	$u_{i',\,i+1}$	$u_{i,\,i+1}+u_{i',\,i+1}$
0100000100000110000000011000001101100011011000001100111000000000		0	11	11
1000000000000000000000000000000000000000010000011100000000000100	1	0	2	2
0000000010001000000001000100000000000000000000000000000000000010	0	0	0	0
0010000100000010000000000001000000000001000100000100010010010000	0	0	2	2
0100010001000000011100000000000001000000000001000000000000000000	0	0	2	2
0000000000000000000000000000000000000000000000000000000000000010	0	0	2	2
0010000011000100000000000000000000000000000000000000000000011000	0	3	1	4
0110000110000110000110000100000000010010010010010010010010010010	0	4	5	9
0101010010010000110000110000110000110000110000000010000000000001	0	3	6	9
1111010000000000000000000010000000000010000010000000000010110111	1	1	7	8
1001000000000000011111010000000000000000000000000000010000000000	1	0	7	7
0000001000100111000000000000101000000000000000010100100001100101O	0	2	5	7
1110010100000000000000010100100001000000000100100000000000000000	1	1	5	6
0100100000100000000000111110000000000000111110000001110101000000	0	4	8	12
1010100000000000101010000000100000000000100010100000000010100010	1	8	4	12
0000000000000000010001001000100010011011001110110110100000100010	0	4	0	4
0010101010001000100000000000000000010001000100100100010001000000	0	4	11	15
1000000000000111000001111100000111000000111110100000101010000001	1	4	12	16
0100000100010000010000000000010000000100001110000100000100000011000	0	5	1	6
0000001000001000100010000100000100001100001000001000010001000100	0	4	4	8
0100000100000110000000011000001101100011011000001100111000000000	0	—	—	—
	U_{line}	47	U_{column} 95	$U=142$

For fairly large images, it is difficult to try all the possible alternative line lengths, especially in manual work. We will therefore propose a less reliable, and yet much faster method. The same method has been applied for image decoding on a screen of an arbitrary shape, when it is in principle impossible to examine all the alternatives.

If a boundary element is interposed between two elements, its position from some initial point μ can be identified as i. We will say that the point μ_i has a U-neighborhood if on both sides of the boundary symbol in position i from the origin there are segments d such that when the right-hand segment is placed under the left-hand segment, we obtain two lines λ_1, λ_2 for which $u_{1,\,2}+u'_{1,\,2}\geqslant U$.

Let the set of points with a U-neighborhood be $\{\mu_i(U)\}$. The simplest procedure using the properties of U-neighborhoods is based on the assumption that there exists $U_{\max}$ such that $\{\mu_i(U_{\max})\}=1$ (the power of the set of points with a U-neighborhood is 1). This is interpreted as follows: the image contains a pair of best lines, which are very close to each other and pass through the part of the image rich in details. Thus, $\{\mu_i(U_{\max})\}$ contains a single point μ_{best}. The length of the U-neighborhood for μ_{best} is the length of the line; μ_{best} itself is the reference point.

The following procedure therefore can be applied: setting $U=1, 2, 3, \ldots$ we establish whether or not a particular value of U is attained for more than

a single point. If this is so, U is increased by unity, and the search is repeated. Otherwise, we have located a single point with a U-neighborhood. Given this point (μ_{best}), we use the length of its neighborhood to determine the length of the line; the position of the point identifies the beginning (the end) of the lines. If there is no such point, the line length is identified with the length of the U-neighborhood of one of the points with a $(U-1)$-neighborhood. In this case, the solution is not single-valued.

Example. Decode the pattern

0101001010011100001000001.

1) Is there more than one point μ_i for which $U \geqslant 1$? The answer is yes, e.g., μ_2, μ_7.

2) Is there more than one point for which $U \geqslant 2$? Again yes, e.g., the same points μ_2, μ_7.

3) Is there more than one point for which $U \geqslant 3$? No, there is one point μ_5 for which $U_{max} = 4$; the length of the line is 5.

The answer to the decoding problem is the outline of the numeral 4 on a rectangular screen.

When decoding patterns on a screen of arbitrary shape, μ_{best} is obtained according to the same rules; one line is then added from above and one from below to the selected pair of lines, whose length and position are chosen to ensure a maximum increment in U. Changing the beginning and the end of the adjacent lines, without altering the value of U, we ensure maximum smoothness of the boundaries, using one of the proposed functions.

Example. Consider the message

0000001111100000000000100010000000001111100000001010100000
10101000111110000000000.

All the points μ_{best} lie densely between position 53 and position 58 (this is a property of messages on an arbitrary screen). The pair of lines corresponding to one of the points is given below:

0000010101
0000010101

For these lines $U_{1,2} = 6$.

The best position of the next line is the following:

0000010101
0000010101
00011111

Here $U_{2,3} = 2$.

Construction of the next lines does not alter the value of U. The lines stacked on top are

00000011111
0000000000010001
0000000000111110
00000010101

and on the whole

```
      00000011111
000000000010001
 000000000111110
    0000010101
    0000010101
     00011111
```

which gives the correct answer: the pattern of a "window." The original pattern is

```
00000011111000000
000001000100000
0000111110000
00010101000
001010100
0111110
00000
000
0
```

i.e., a "window" on a triangular screen. Thus, although knowledge of the screen geometry is essential, it does not alter the pattern itself.

§8. ALGORITHMS ANALOGOUS TO ALGORITHMS WHICH CONSTRUCT BILINGUAL DICTIONARIES

Letter-comparison algorithms using the properties of close neighborhoods

In the previous sections we described examples of various algorithms analyzing texts written in an unknown language.

We have mentioned before that the aim of this analysis is to construct the best interpretation containing information needed for the most effective forecasting of the inaccessible part of the text for any arbitrary partition of the text into accessible and inaccessible parts.

Suppose that such an interpretation has been found. Examining the accessible part of the text, we will be in a position to predict what comes next. We will possibly learn to construct "correct sentences" or even "correct texts" in the new language.

The next question, however, is concerned with a more fundamental aspect: is this really what we sought to achieve when we started the decoding? The answer is an emphatic no. After all, we still do not know how to translate the message into a known language and into the "language of reality." In other words, to give a crude example, we still cannot build the machine that the message describes.

To effectively translate an unknown text, we should establish a correspondence between the elements of our language and some elements of the

code message. The corresponding elements in either language may be selected by a variety of techniques; for example, we can compile a list of sentences in the unknown language and their translations into our language; or we may assemble lists of words with the appropriate word translations. The best and the most natural approach is probably to compare certain linguistic phenomena on which the predicate system of the two languages is based. We have mentioned previously that the basis for the analysis of textual meaning is provided by the "semantic classes of words." A bilingual dictionary with ordinary words replaced by names of semantic classes would be shorter and better than a conventional bilingual dictionary; a dictionary of sentences, on the other hand, is impracticable and cannot be drawn up even for a pair of known languages.

Translation from one language into another thus requires bilingual dictionaries of certain elementary phenomena which make the text. This is a necessary condition, but obviously insufficient. We should, moreover, be able to compare the rules according to which the elements of the two languages combine between themselves. After all, the same words can be used to give sentences with entirely different meanings.

In other words, we should be able to define the "closeness relation" in the two languages and, when preparing the translation, we have to ensure that the words of the translation are represented by the same closeness relations as the words of the source text.

The decoding algorithm, however, is never expected to produce a polished and styled translation. It is quite enough if the algorithm provides sufficient information for a human operator to prepare the finished translation. The development of a "dictionary" and "comparative grammar" is therefore one of the last aims of decoding algorithms.

We will describe an algorithm which compiles a dictionary of sorts, but the component elements of this dictionary are letters, rather than words or semantic units. The starting assumption is that a certain "anthropomorphic" (i.e., vocal) language is to be decoded and translated into another known human language. We know how the letters of the known language are pronounced, but the pronunciation of the letters in the other language is unknown. Our aim is to describe the pronunciation of the letters of the unknown language using the letters of the known language. In the simplest case, this can be achieved by establishing a "correct" one-to-one correspondence between the letters of the unknown language and those of the known language.

We will describe a simple algorithm which establishes this correspondence or substitution. We will also consider certain means for finding more general solutions.

The set of permissible solutions in this particular case is the set of all one-to-one mappings of the letters of the unknown language into the letters of the known language. If the two alphabets are of unequal length, the smaller of the two should be supplemented with an appropriate number of letters which are interpreted as "accidentally missing" from the text.

The basic hypothesis used in the construction of the quality criterion for the letter substitutions is that letters conveying similar sounds should have similar "combination properties." A particular form of the algorithm depends on the precise interpretation of the concept of "combination properties."

A similar assumption is naturally introduced when compiling dictionaries of "semantic multipliers"* and words. It is assumed that elements of similar significance or meaning follow the same combination pattern. Although in certain cases this assumption is not absolutely obvious, it nevertheless provides the only conceivable basis for decoding.

In this algorithm, the combination properties of the letters will be represented by a table T of the frequency of occurrence of the various letter pairs. The number $P(a_i, a_j)$ at the intersection of row i and column j is expressed by the equality $P(a_i, a_j) = \frac{\varphi(a_i a_j) + \varphi(a_j a_i)}{2N}$, where a_i and a_j are the letters of the given alphabet, $\varphi(a_i, a_j)$ is the number of times the pair of letters a_i and a_j occurs in the text, N is the length of the text.

The combination properties of a letter a_i are defined by the row of numbers of the form $P(a_i, a_x)$ in the table T.

The dissimilarity of two letters a_i and a_j in the same language can be measured using one of the equations expressing distance between the points of an n-dimensional space.

We define the following measure of dissimilarity of the letters a_i and a_j:

$$\bar{\sigma}(a_i, a_j) = \sum_{k=1}^{n} |P(a_i, a_k) - P(a_j, a_k)|,$$

where a_i, a_j, a_k are letters of the given alphabet.

Now consider a certain substitution π: $A \to B$ of the unknown alphabet A into the known alphabet B. Let a'_i be the image of the letter a_i under the substitution π. Then the function

$$\bar{\sigma}(a_i, a'_i, \pi) = \sum_{k=1}^{n} |P(a_i, a_k) - P(a'_i, a_k)|$$

characterizes the dissimilarity (with regard to combination properties) of the letter a_i and its image a'_i under the substitution π.

The quality of the substitution π can be found as the sum of dissimilarities of the letter pairs entering this substitution, provided their dissimilarity is evaluated for the same substitution. For the quality ρ of the substitution we thus have

$$\rho = \sum_{i=1}^{n} \bar{\sigma}(a_i, a'_i, \pi). \qquad (Z)$$

We should thus look for the best substitution π_{best}, such that

$$\rho(\pi_{best}) = \min.$$

In principle, π_{best} could be found by examining all the possible substitutions and estimating the quality ρ of each using equation (Z). However, the number of possible substitutions is very large ($n!$), and we will therefore propose an abbreviated procedure which, we hope, is equivalent to the complete procedure for all practical purposes.

We shall say that π_j is a transposition of π_i, or π_j (π_i), if to some pairs $a_r \to a'_r$, $a_s \to a'_s$ of the substitution π_i correspond the pairs $a_r \to a'_s$, $a_s \to a'_r$ of the substitution π_j, whereas the other pairs of the two substitutions coincide.

* I.e., names of classes of semantically close words.

As is known, for any two substitutions π_u and π_w, we can always construct a sequence of substitutions $\pi_x, \ldots, \pi_i, \ldots, \pi_1$, where π_1 is a transposition of π_w; π_x is a transposition of π_u, and for any pair π_i $\pi_{i-1} (i \leqslant x;\ i > 1)$, π_i is a transposition of π_{i-1}. The sought substitution is therefore always attainable by advancing from any given substitution through a number of intermediate substitutions, as indicated.

The transposition π_j of some substitution π_i may be a best substitution only if $\rho(\pi_j) \leqslant \rho(\pi_i))$. If for some π_i and any $\pi_j (\pi_i)$ we have $\rho(\pi_j) \geqslant \rho(\pi_i)$, we say that the function ρ has a local minimum at the point π_i. Clearly, if ρ has an absolute minimum at the point π_{best}, it also has a local minimum at the same point π_{best}.

Therefore we can find the absolute minimum π_{best} by examining all the possible local minima.

We cannot propose a suitable method at this stage, but we will describe a procedure which locates a sufficiently deep local minimum.

For some substitution π_i, we define the set of transpositions $\{\pi_j\ (\pi_i)\}$. For each transposition $\pi_j\ (\pi_i)$ we calculate the increment $\Delta\rho$, equal to $\rho(\pi_i) - \rho(\pi_j)$. Then we choose π_q such that $\Delta\rho(\pi_q)$ is maximum and positive in the set of the increments $\{\Delta\rho(\pi_j(\pi_i))\}$. A similar procedure is repeated for $\{\pi_t(\pi_q)\}$. The routine ends when we have found a substitution for which no transpositions with positive increments exist.

To calculate the increment $\Delta\rho$, we have to determine $\rho(\pi_i)$ and $\rho(\pi_j)$. If π_i differs from π_j in that π_i contains the pairs $a_r \to a'_r, a_s \to a'_s$ and π_j the pairs $a_r \to a'_s$, $a_s \to a'_r$, the increment $\rho(\pi_j) - \rho(\pi_i)$ can be calculated using the following somewhat cumbersome formula:

$$\begin{aligned} &2\sum_x |P(a_r, a_x) - P(a'_r, a'_x)| + 2\sum |P(a_s, a_x) - P(a'_s, a'_x)| - \\ &- |P(a_r, a_r) - P(a'_r, a'_r)| - |P(a_s, a_s) - P(a'_s, a'_s)| - \\ &- 2\sum_x |P(a_r, a_x) - P(a'_s, a'_x)| - 2\sum_x |P(a_s, a_x) - P(a'_r, a'_x)| + \\ &+ |P(a_r, a_r) - P(a'_s, a'_s)| - |P(a_s, a_s) - P(a'_r, a'_r)|. \end{aligned}$$

As the initial substitution, we can choose the one that is obtained when the two alphabets are arranged one next to the other in the order of decreasing numbers,

$$\sum_x P(a_i, a_x) \quad \text{and} \quad \sum_x P(a'_i, a'_x),$$

respectively.

The initial substitution will include pairs of letters $a_i \to a'_i$ which have the same number i in the corresponding sequences.

We will now summarize the algorithm in the form of a system of generalized instructions.

1. For the text to be translated, draw up a table T_1 of numbers

$$P(a_i, a_j) = \frac{\varphi(a_i a_j) + \varphi(a_j a_i)}{2N},$$

where $f(a_i a_j)$ is the number of times the pair $a_i a_j$ occurs in the text, N is the length of the text.

2. Draw up a similar table T_2 for the text in the known language.

3. Is the number of rows in T_1 and T_2 the same? If not, complete the smaller table with dummy rows consisting of zeros until the two tables are of the same size.

4. Arrange the rows and the columns of the two tables in the order of decreasing numbers $\sum_x P(a_i, a_x)$ and $\sum_x P(a'_i, a'_x)$.

5. Construct all the $\frac{n(n-1)}{2}$ possible pairs of the form a'_i, a'_j and for each pair calculate the increment $\Delta\rho(a'_i, a'_j) = \Delta\rho$.

6. Are there any positive $\Delta\rho(a'_i, a'_j)$? If none, proceed to instruction 7; otherwise, go on to instruction 9.

7. Print out the answer: the set of pairs $a_i \rightarrow a'_i$.

8. End.

9. Interchange row and column a'_i with row and column a'_j in the second table for which $\Delta\rho(a'_i, a'_j)$ is maximum; return to 5.

In the case of combination properties represented by an asymmetric matrix with entries of the form $P(a_i\ a_j) = \frac{f(a_i\ a_j)}{N}$, we should proceed along the same lines. The only difference in this case is that the increments Δ have to be calculated using a different formula.

Let us now briefly describe an algorithm using a more general transformation.

We are looking for a transformation Δ which meets the following requirements.

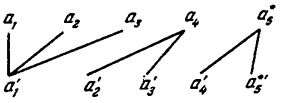
1. For any letter a_i $(a_i \in A)$, there is a pair $a_i a'_i \in \Pi$.

2. For any letter a'_j $(a'_j \in B)$, there is a pair $a_j, a'_j \in \Pi$.

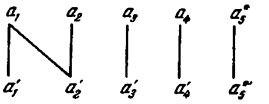
3. To each of the alphabets, at least one "dummy" letter a^*_i is added, such that $p(a^*_i, a_x) = 0$ for any a_x.

4. When conditions 1—3 are observed, the sum of the numbers $P(a_i, a'_i)$ is minimum.

Thus, a mapping of the form



is permissible, while the mapping



is unacceptable: here the pair $a_1 a'_2$ can be omitted without breaking conditions 1 through 3.

The distinctive feature of an algorithm using this mapping is that the length of each row in tables T_1 and T_2 is doubled, and this in addition to the elementary manipulations which interchange the rows of the second table.

The elementary manipulations can be divided into induced and free. Indeed, the result of some manipulation may be an unacceptable mapping. To ensure an acceptable mapping, some additional elementary manipulations should be carried out.

Free elementary manipulations are those which are not intended to restore the mapping to an acceptable form. The true gain of an elementary manipulation will be defined as the gain of the elementary manipulation proper plus the sum of the gains of the best elementary manipulations induced by the particular elementary manipulation and restoring the mapping to an acceptable form.

At every step of the routine, we have to calculate the true gain of all the free elementary manipulations. Then if the gain is positive, we have to carry out the entire sequence of elementary manipulations corresponding to the truly best free elementary manipulation. The routine is terminated when the true gains of all the free elementary manipulations become negative.

Computer experiments were carried out using this algorithm. English and French texts of 5000 letters each were selected for this purpose.

The correspondence is shown in Figure 60.

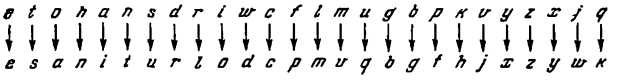


FIGURE 60. A substitution obtained in a computer experiment analyzing the correspondence of English and French letters.

Tables 4.8 and 4.9 contain the relative frequencies of two letter sequences in English and French. For convenience, all the numbers have been multiplied by a factor of 10,000. The order of letters in the corresponding sequences was ignored.

The rows (and columns) with numbers of the same order of magnitude should be close to one another, since the tables have been processed by the algorithm. We must confess that the similarity between the corresponding rows is not very striking.

The results of the comparison are evidently quite unsatisfactory. In any case, they are no better than what could be obtained by a simple frequency analysis of letters.

Let us try to establish the reasons for the poor correspondence.

In our opinion, the main reason is to be sought in distortions of orthographic origin. Thus, the two-letter combination "th" conveys a single sound in English; the letter "y" in English is sometimes pronounced as a vowel and sometimes as a consonant. It is surprising, however, that the same effect has not distorted the vowel and consonant identification procedure.

The inadequate results of the computer experiment focus our attention on the tremendous difficulties in the direct comparison of the linguistic elements of different languages.

We will now show that a preliminary analysis markedly improves the quality of the comparison procedure. Let us first divide the letters of both languages into vowels and consonants (this algorithm, as we know, yields virtually error-free results), and then compare tables corresponding to the two classes V (vowels) and C (consonants).

TABLE 4.8. English

	e	t	o	h	a	n	s	d	r	i	w	c	f	l	m	u	g	b	p	k	v	y	z	x	j	q
e	440	774	160	2173	574	794	1059	1059	1484	230	530	420	265	605	515	80	220	330	270	170	345	230	250	75	15	5
t	774	520	749	2018	804	530	704	295	325	759	175	80	245	125	95	315	80	40	55	35	0	120	5	20	5	0
o	160	749	440	305	85	624	350	270	654	195	385	205	639	405	315	410	110	175	140	130	75	100	10	0	5	0
h	2173	2018	305	70	659	120	385	215	145	530	235	215	105	40	5	60	260	10	25	30	5	45	0	0	0	0
a	574	804	85	659	20	1054	829	505	644	224	589	340	165	455	325	95	190	110	290	55	130	195	10	0	10	0
n	794	530	624	120	1054	10	250	819	125	1109	110	135	35	35	15	175	530	35	15	110	30	40	10	5	0	0
s	1059	704	350	385	829	250	500	200	220	709	190	145	140	140	140	280	50	60	130	75	25	90	0	0	10	5
d	1059	295	270	215	505	819	200	90	145	355	90	35	65	215	70	40	35	125	30	20	20	40	0	0	0	0
r	1484	325	654	145	644	125	220	145	120	395	60	90	150	75	85	290	70	115	120	15	15	90	0	0	5	0
i	230	749	195	530	225	1109	709	355	395	0	215	215	155	435	270	40	300	40	125	95	135	40	15	35	0	0
w	530	175	385	235	589	110	190	90	60	215	60	0	40	30	20	5	5	0	5	65	0	50	0	0	0	5
c	420	80	205	215	340	135	145	35	90	215	0	30	20	85	5	80	25	5	5	60	0	15	0	5	0	0
f	265	245	640	105	165	35	140	65	150	155	40	20	130	140	20	50	20	25	10	0	10	50	0	0	0	5
l	604	125	405	40	455	35	140	215	75	435	30	85	140	490	20	180	105	105	60	70	5	195	5	0	0	0
m	515	95	315	5	325	15	140	70	85	260	20	5	20	20	30	95	20	15	70	0	0	50	0	0	0	0
u	75	315	410	60	95	175	280	40	290	40	5	80	50	180	95	0	110	85	70	5	0	15	5	0	20	25
g	230	80	110	260	190	530	50	35	70	300	10	25	20	105	20	110	40	15	5	10	5	20	0	0	0	0
b	330	40	175	10	110	35	60	125	115	40	0	5	25	105	15	85	15	10	0	0	0	80	0	0	5	0
p	270	55	140	25	290	15	130	30	120	125	5	5	10	60	70	70	5	0	140	5	0	10	0	20	0	0
k	170	35	130	30	55	110	75	20	15	95	65	60	0	70	0	5	10	0	5	0	0	10	0	0	0	0
v	345	0	75	5	130	30	25	20	15	135	0	0	10	5	0	0	5	0	0	0	0	10	0	0	0	0
y	230	120	100	45	195	40	90	40	90	40	50	15	50	195	50	15	20	80	10	10	10	0	0	0	5	0
z	25	5	10	0	10	10	0	0	0	15	0	0	0	5	0	5	0	0	0	0	0	0	5	0	0	0
x	75	20	0	0	0	5	0	0	0	35	0	5	0	0	0	0	0	0	20	0	0	0	0	0	0	0
j	15	5	5	0	10	0	10	0	5	0	0	0	0	0	0	20	0	5	0	0	0	5	0	0	0	0
q	5	0	0	0	0	0	5	0	0	0	5	5	5	0	0	25	0	0	0	0	0	0	0	0	0	0

TABLE 4.9. *French*

	e	s	a	n	i	t	u	r	l	o	d	c	p	m	v	q	b	g	f	h	j	x	z	y	w	k
e	241	2063	280	1758	624	1344	1004	1698	1778	75	1129	949	574	884	654	115	160	215	175	190	175	85	165	40	10	0
s	2063	1169	934	435	849	564	619	315	370	345	240	185	145	95	75	80	60	35	70	15	45	10	5	20	0	0
a	280	934	50	684	924	579	380	684	1149	15	300	270	514	415	290	25	145	155	180	135	55	30	10	80	0	0
n	1758	435	684	400	525	649	574	85	110	889	315	225	65	400	100	40	15	145	50	25	30	5	0	10	5	0
i	624	849	924	524	20	1064	250	579	634	405	305	70	65	175	140	90	100	95	85	65	5	85	15	0	0	0
t	1344	565	579	649	1064	270	485	520	300	330	180	195	255	50	15	60	30	15	20	0	25	25	5	15	0	0
u	1004	619	380	574	250	485	40	679	280	929	300	185	80	100	105	575	70	40	30	5	85	80	5	5	0	0
r	1698	315	684	85	579	520	679	140	210	425	200	205	205	80	95	50	110	90	35	0	15	10	0	40	0	0
l	1778	370	1149	110	634	300	280	210	470	300	40	50	175	15	15	65	85	10	60	15	5	25	20	35	0	0
o	80	325	15	889	405	330	929	425	300	0	80	450	245	305	285	5	240	55	85	70	60	10	5	25	5	0
d	1129	240	310	315	305	180	300	200	40	80	50	0	15	15	0	5	0	5	10	25	0	20	30	5	0	0
c	999	185	270	225	65	195	185	205	50	450	0	10	15	15	5	15	0	10	0	229	0	30	15	5	0	0
p	575	145	515	65	65	255	80	205	175	245	15	15	70	145	0	0	0	0	0	30	0	15	5	0	0	0
m	884	95	415	40	175	50	100	75	15	305	15	15	145	145	0	5	65	10	0	0	0	10	20	10	0	0
v	654	75	290	100	140	15	005	95	15	285	0	5	0	0	0	0	0	0	0	0	0	10	25	5	0	0
q	115	80	25	40	90	60	574	50	65	5	5	15	0	5	0	0	0	9	0	0	0	09	10	0	0	0
b	160	60	145	15	100	30	70	110	85	240	0	0	0	65	0	0	0	0	0	0	0	5	0	0	0	0
g	215	35	155	145	95	15	40	90	10	55	5	5	0	10	0	5	0	5	0	0	0	15	0	0	0	0
f	175	65	180	45	85	20	30	35	60	85	10	0	0	0	0	0	0	0	90	0	0	0	0	0	0	0
h	190	15	135	25	65	0	5	0	15	70	35	220	30	0	0	0	0	0	0	0	0	0	5	0	0	0
j	175	45	55	25	5	25	85	15	5	60	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
x	85	10	30	5	85	25	80	10	25	10	20	30	15	10	10	5	5	15	0	0	5	0	0	0	0	0
z	165	5	10	0	15	5	5	0	20	5	30	15	5	20	25	10	0	0	0	5	0	0	0	0	0	0
y	40	20	75	10	0	15	5	40	35	25	5	5	0	10	5	0	0	0	0	0	0	0	0	0	0	0
w	10	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	V	C
V		
C		

Since the numerical entries in the off-diagonal blocks VC and CV are substantially greater than those in the diagonal blocks VV and CC, it suffices to calculate the sums in VV and CC. We have*

	VV	CC
French text	1972	4302
English text	1330	5261

We conclude that the first class of French letters (vowels) corresponds to the first class of English letters. We further divide the vowels of each language into subclasses:

French text:

	e, a, o	*i, u, y*
e, a, o	256	785
i, u, y	785	146

English text:

	e, o	*a, i, u, y*
e, o	326	385
a, i, u, y	385	234

A comparison of these tables establishes a correspondence between the French *e, a, o* and the English *e, o*, the French *i, u, y* and the English *a, i, u, y*. Corresponding subdivision of the consonants gives for the French text

s, r, n, l, m, h, x, z (class 1)
b, d, g, p, t, k, c, q, f, v, w, j (class 2)

and for the English text

s, r, n, l, m, h, x, q, j (class 1)
b, d, g, p, t, k, c, f, w, v (class 2)

The diagonal squares of the corresponding tables are the following:

	cl. 1 cl. 1	cl. 2 cl. 2
French	1174	416
English	936	372

We have again obtained a correct correspondence. Further subdivision into smaller categories in principle could provide detailed information on the correspondence of the individual letters of the two languages, but unfortunately,

* The calculations were carried out using Table 4.3.

for fairly small texts, the information concerning combination and frequency properties is too scanty.

The next algorithm possibly can be applied using the preliminary division into classes and restricting the comparison to letters which belong to the corresponding classes only.

The above calculations show that in order to establish a valid correspondence, the "predictive systems" of the text should first be analyzed. In particular, it is clear beyond all doubt that straightforward comparison of letters in the two languages is doomed to failure. It is the "semantic classes" obtained by a separate treatment that should be compared.

An algorithm using distant neighborhoods

The previous algorithms represent the so-called "statistical" approach of the military deciphering techniques. There are algorithms, however, which utilize the alternative conception, namely the method of characteristic words. Let us consider the set of permissible solutions and the quality function of such an algorithm.

This algorithm also establishes a correspondence expressed as a certain substitution. Let us first estimate the quality of the pair a_x, a'_x entering the substitution π. We will designate the letter a_x which occurs in position p in the unknown text by the symbol $a_x(p)$, and the letter a'_x which occurs in position t in the unknown text by the symbol $a'_x(t)$. A pair of letter sequences will be called a permissible u, v-neighborhood of the pair $a_x(p)$, $a'_x(t)$ if for every $k (k < u)$ and $m (m \geqslant v)$ the pair $a_y(p+k), a'_y(t+k)$ and the pair $a_y(p-m)$, $a'_y(t-m)$ are elements of π. The difference $u - v$ will be called the length of the permissible u, v-neighborhood. The maximum length among all the permissible u, v-neighborhoods of the pair a_x, a'_x will be used as the quality of the pair a_x, a'_x, or $q(a_x, a'_x)$. The quality Q of the substitution π can be defined as the sum of the qualities of the constituent pairs:

$$Q(\pi) = \sum_x q(a_x, a'_x).$$

The higher $Q(\pi)$ the better is the substitution. For the best substitution π_{best}, $Q(\pi_{best})$ is minimum.

While the previous algorithms of this section used the properties of nearest neighborhoods, the present algorithm is based on the similarity of long letter sequences. Both principles naturally can be combined into a single procedure.

§9. CLASSIFICATION ALGORITHMS (END)

"Mathematically" correct algorithm for vowel-and-consonant identification

We will describe one of the "mathematically" correct algorithms minimizing the function K_1.

We will say that K_1 has a local minimum for a partition $R=k_1, k_2$ of the alphabet A into two classes, if a transfer of any letter from the class k_2 into the other class k_1 and vice versa does not generate a partition k_1^*, k_2^* for which K_1 is smaller than for $R=k_1, k_2$.

If there is no partition such that K_1 is smaller than for R, we say that the function K_1 has an absolute minimum for the partition R.

Clearly, if the function has an absolute minimum for the partition R, it also has a local minimum for that partition. Therefore, the absolute minimum can be found by examining all the local minima (i.e., all the partitions for which K_1 has a local minimum).

A subset $\mathfrak{K}$ ($\mathfrak{K} \subset A$) is said to be permissible if for every a_l ($a_l \in \mathfrak{K}$) we have

$$\sum_{l=m+1}^{n} \varphi(a_l, a_k) - \sum_{i=1}^{m} \varphi(a_l, a_j) \geqslant 0.$$

Here $m = |\mathfrak{K}|$, $a_l \in \mathfrak{K}$; $a_j \in \mathfrak{K}$; $a_k \in A \setminus \mathfrak{K}$.

The following two theorems hold true:

Theorem 1. If both classes of partition k_1, k_2 are permissible, the function K_1 has a local minimum for k_1, k_2.

Suppose that the proposition of the theorem is not true, i.e., there exists a letter a_x which reduces the value of K_1 when transferred from k_1 into k_2. The classification obtained when a_x is transferred from k_1 into k_2 is designated k_1', k_2', so that $k_1', k_2' = k_1 \setminus a_x, k_2 \cup a_x$. The value of K_1 for k_1', k_2' is

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \varphi(a_i, a_j) + \sum_{k=m+1}^{n}\sum_{l=m+1}^{n} \varphi(a_k, a_l) - 2\sum_{l=1}^{m} \varphi(a_x, a_l) + 2\sum_{k=m+1}^{n} \varphi(a_x, a_k).$$

The value of K_1 for k_1, k_2 is

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \varphi(a_i, a_j) + \sum_{k=m+1}^{n}\sum_{l=m+1}^{n} \varphi(a_k, a_l).$$

By assumption, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \varphi(a_i, a_j) + \sum_{k=m+1}^{n}\sum_{l=m+1}^{n} \varphi(a_k, a_l) - \sum_{i=1}^{m}\sum_{j=1}^{m} \varphi(a_l, a_j) - \sum_{k=m+1}^{n}\sum_{l=m+1}^{n} \varphi(a_k, a_l) + 2\sum_{i=1}^{m} \varphi(a_x, a_i) - 2\sum_{k=m+1}^{n} \varphi(a_x, a_k) > 0$$

or

$$\sum_{i=1}^{m} \varphi(a_x, a_i) - \sum_{k=m+1}^{n} \varphi(a_x, a_k) > 0,$$

which is impossible, since a_x belongs to a permissible class and thus satisfies the inequality

$$\sum_{k=m+1}^{n} \varphi(a_x, a_k) - \sum_{i=1}^{m} \varphi(a_x, a_i) \geqslant 0.$$

Theorem 2. If the subset $\mathfrak{K}_r$ is impermissible, any subset $\mathfrak{K}_t$ containing $\mathfrak{K}_r$ is also impermissible.

Since $\mathfrak{K}_r$ is impermissible, there exists a letter a_x, such that

$$\sum_{k=m+1}^{n} \varphi(a_x, a_k) - \sum_{i=1}^{m} \varphi(a_x, a_i) < 0,$$

where a_i, a_x are elements of $\mathfrak{K}_r$, and a_h is not an element of this subset.

Consider the following expression:

$$\sum_{r=l+1}^{n} \varphi(a_x, a_r) - \sum_{s=1}^{l} \varphi(a_x, a_s),$$

where $|\mathfrak{K}_t| = l$; a_x is naturally an element of $\mathfrak{K}_t$, $a_s \in \mathfrak{K}_t$, $a_r \in A \setminus \mathfrak{K}_t$. It is readily seen that

$$\sum_{r=l+1}^{n} \varphi(a_x, a_r) = \sum_{k=m+1}^{n} \varphi(a_x, a_k) - \sum_{p=m+1}^{l} \varphi(a_x, a_p),$$

where $a_p \in \mathfrak{K}_t \setminus \mathfrak{K}_r$, and hence

$$\sum_{r=l+1}^{n} \varphi(a_x, a_r) \leqslant \sum_{k=m+1}^{n} \varphi(a_x, a_k);$$

at the same time

$$\sum_{s=1}^{l} \varphi(a_x, a_s) = \sum_{i=1}^{m} \varphi(a_x, a_i) + \sum_{p=m+1}^{l} \varphi(a_x, a_p),$$

so that $\sum_{s=1}^{l} \varphi(a_x, a_s)$ is greater than or equal to $\sum_{i=1}^{m} \varphi(a_x, a_i)$. All the more so,

$$\sum_{r=l+1}^{n} \varphi(a_x, a_r) - \sum_{s=1}^{l} \varphi(a_x, a_s) \leqslant \leqslant \sum_{k=m+1}^{n} \varphi(a_x, a_k) - \sum_{i=1}^{m} \varphi(a_x, a_i),$$

and it is thus negative.

The two theorems lead to a construction of an algorithm which examines all the permissible classes. Theorem 2 enables us to shorten the examination procedure. We then seek all the possible nonintersecting pairs of permissible classes which cover the entire alphabet. Every one of these pairs is a local minimum in virtue of Theorem 1. The absolute minimum is picked out from among the local minima by direct examination of the alternatives. Let us summarize this algorithm in instruction form:

1. Construct all the possible subsets of the alphabet, containing t letters each.
2. Omit those subsets which are not permissible.
3. Draw up a list of all subsets containing $t + 1$ letters each and not containing any t-letter impermissible subsets.

4. Is this list empty? Yes: proceed to instruction 5. No: substitute $t+1$ for t and return to instruction 2.

5. Find all pairs of nonintersecting permissible subsets covering the entire alphabet, compute the value of K_t for each pair, choose the pair for which this value is minimum; the particular pair of classes provides the solution.

6. End.

The initial value of the parameter t is 2. Theorem 2 helps us with instruction 3 of the routine: in virtue of the theorem, we do not have to construct all the possible subsets containing $t+1$ letters each, but only those which are made up of permissible t-letter subsets. The permissibility check is based on the inequality

$$\sum_{k=m+1}^{n} \varphi(a_x, a_k) - \sum_{i=1}^{m} \varphi(a_x, a_i) \geqslant 0,$$

which should hold true for every letter of a permissible subset.

An algorithm translating syllabic writing into alphabet writing

The vowel-and-consonant identification algorithm assigns the value of one binary feature to each textual element. The algorithm discussed in this subsection, on the other hand, assigns the values of two features to each element, and these are moreover multidigit, and not binary, features.

This algorithm is evidently far from being able to determine on its own the number of distinctive features and their possible values. It nevertheless has a two-fold practical importance. First, it may help to establish the pronunciation of letters in the so-called syllabic writing, and second, an analogous algorithm will be useful for morpheme identification.

There are many examples of writing in which a single element corresponds to a sequence of sounds, rather than to a single sound. When a syllabic text of this kind is to be decoded, it should first be transcribed into normal alphabet writing. Then the pronunciation and the grammar are easier to determine.

The sequence of sounds corresponding to a syllabic element is often not a true syllable. It commonly has the following standard structure: the first sound is consonantal and the second vocalic.

This, in particular, is the case in Creto-Mycenaean writing. For example

In decoding such syllabic writing, the algorithm should translate each syllabic symbol into a pair of symbols, whereby the second symbol is common for all syllabics with the same main vocalic sound, and the first symbol is common for all the syllabics with the same consonantal sound.

The set of these symbols constitutes an "ordinary" alphabet replacing the syllabary.

The morphological interpretation of the algorithm will be discussed later on.

The algorithm is built as follows. Two classifications k_1 and k_2 of the syllabary S are built. According to the first classification, syllabics with a common vocalic part, e.g., *ta*, *pa*, *ka*, etc. are grouped in one class; according to the second classification syllabics with a common consonantal part (*ta*, *tu*, *to*, ..., etc.) are in one class.

The classes from k_1 are identified by the symbols $\alpha_1, \alpha_2, \ldots$, and the classes from k_2 by the symbols $\beta_1, \beta_2, \ldots$. If k_1 is a vocalic classification, each syllabic can be assigned a sequence of the form β_i, α_j. An ordinary alphabet A corresponding to the syllabary comprises the symbols $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \ldots, \beta_k$. The alphabet A may also contain symbols for "null" vowels and consonants. A "null vowel," i.e., a symbol denoting the absence of a vowel, is required to transcribe a syllabic which corresponds to an unpaired constant, and a "null consonant" is required for the transcription of vowel syllabics.

It turns out that k_1 and k_2 cannot be defined arbitrarily. The sought pair of classifications should have a certain restrictive property. If a given pair has this property, we will refer to it as a permissible pair.

Having defined the set of permissible pairs, we examine it for the best pair, i.e., the one extremizing some quality function.

Let us describe a permissible pair of classifications. It is readily seen that if one classification contains m classes, each class of the other classification will contain at most m elements. Suppose that k_1 comprises two classes, the class of syllabics corresponding to the vocalic sound a and the class of syllabics corresponding to the vocalic sound i. Consider the class of syllabics in k_2 which contain the consonant p, say. It will naturally contain the two symbols *pa* and *pi*, or only one of these syllabics (since some elements of the syllabary S may be missing from the text). No third element in addition to *pa* and *pi* may enter this class, since there is no third vowel to combine with the characteristic consonant of the class.

In some syllabic languages, different syllabics may convey the same sound sequence, but then our algorithm is inapplicable.

It is moreover assumed that pairs of classifications for which fewer syllabics are missing from the text are relatively more likely to be valid.

Let us consider the quality criterion to be applied to a permissible pair. Consider a table $T_1 = \| f(s_i s_j) \|$ where the entry in row i and column j is the number of ordered groups of the form $s_i s_j$ occurring in the text. The row i of the table therefore contains numbers which indicate which elements of S follow some given s_i and with what frequency. Let s_i be decoded as β_k, α_l. It would naturally seem that the appearance of the row i depends to a greater extent on α_l than on β_k, and therefore rows corresponding to combinations ending with α_l should be close to one another. Conversely, close columns are those which correspond to groups beginning with the same letter.

Closeness of rows and columns can be estimated in terms of some distance between the points of an n-dimensional space (where $n = |S|$). We will use the simple relations

$$\bar{\sigma}_1(s_i, s_j) = \sum_{k=1}^{n} |\varphi(s_i s_k) - \varphi(s_j s_k)| \quad \text{for rows, and}$$

$$\bar{\sigma}_2(s_i, s_j) = \sum_{k=1}^{n} |\varphi(s_k s_i) - \varphi(s_k s_j)| \text{ for columns.}$$

Here $\bar{\sigma}_1$ and $\bar{\sigma}_2$ indicate dissimilarity (distance), and s_k runs through the entire syllabary.

We can now construct two tables T_2 and T_3 presenting the row and the column distance of the table T_1.

Let the adjacent rows in T_2 and T_3 correspond to the same class, so that the table can be divided into strips, each containing rows of one class only. The tables will take the form shown in Figure 61.

The hatched squares in Figure 61 contain entries which indicate the distance of elements of a certain class from other elements of the same class. In virtue of our basic assumption, these distances should be small, and the sum of the numbers in all these squares will also be relatively small if the classification is likely.

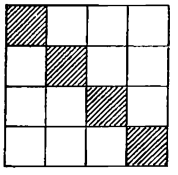


FIGURE 61. Partition of a table into strips corresponding to vocalic classes.

We designate the sum of entries in the quasidiagonal squares of the table T_2 by Σ_1, and the corresponding sum for the table T_3 by Σ_2. To construct a quality function, we use the obvious line of reasoning, which maintains that in a good pair of classifications, even the worst classification is sufficiently good. Therefore, we may take

$$K_4(k_1, k_2) = \max(\Sigma_1, \Sigma_2).$$

Here K_4 is the quality of a permissible pair. Unfortunately, the meaning of K_4 is clear only when the number of elements in each class of one of the classifications is precisely equal to the number of classes of the other classification. In our more general case, we either have to use a more complicated function or to alter the definition of a permissible pair. We adopted the second course.

The final definition of a permissible pair is therefore the following:

1. If one of the classifications contains m classes, the number of elements in any of the classes of the other classification is at most m.

2. The sum of the distances from an element s_i to other elements of the same class is less than the sum of the distances from s_i to elements of any other class.

3. With conditions 1 and 2 satisfied, no other pair of classifications k_1' and k_2' exists which satisfy conditions 1 and 2 and such that

$$|k_1'| = |k_1| \text{ and } |k_2'| < |k_2|.$$

The simplest search algorithm for a permissible pair of classifications calls for a detailed examination of all the possible alternatives, rejecting the impermissible ones and picking out among the remainder the one with minimum K_4. This algorithm, however, is quite impracticable.

We will therefore propose a shorter routine for deriving the best solution which, we hope, is equivalent to the complete examination procedure for all practical purposes.

We construct two sequences of classifications $k_{11}, k_{12}, k_{13}, \ldots$ and $k_{21}, k_{22}, \ldots$. Table T_2 corresponds to one of the sequences, and T_3 to the other.

Using one of the tables, we proceed to construct progressively finer classifications, whereas the other table is used to construct progressively coarser classifications. The sums Σ_{1i} decrease, while the sums Σ_{2i} increase.

The value of K_4 is first determined by the sum Σ_{1i} and then by Σ_{2i}, as shown in Figure 62.

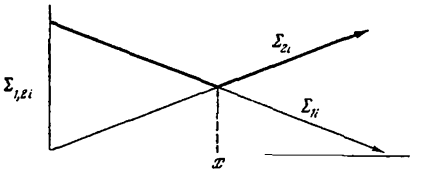


FIGURE 62. The functions $\Sigma_{1,2}$ and K_4.

The thick line in the figure is the plot of K_4. We see from the figure that the minimum of the function occurs at the point x, i.e., the sequence of the classification pairs can be terminated when we reach a pair of classifications built on the basis of rules 1 through 4 in which Σ_{2i} is greater than Σ_{1i}.

We will not go into a more detailed description of the computation procedure. It suffices to note that it largely draws upon the algorithm described on p. 155.

Our algorithm may be given an interesting twist if it is applied to words, rather than syllabics. Each word may be regarded as an element of some "grammatic" classification and an element of some "lexic" classification. For example, the Russian word "домой" [home, in the sense of "going home"] is classified in one grammatic class with the words "столом", "водой", "решением", [declensions of the words "table," "water," "solution"] in that they are all instrumental case singular. On the other hand, this word is part of another class containing the words "дом", "дома", "домах", etc., all based on the same root "дом" [house or home]. The relationship to one of the grammatic classes in Russian is generally identified by the suffix. This is not always so, however, and the nominative case has no identifying suffixes. In this case we are dealing with "null suffixes."

The entire situation is completely analogous to the various assumptions regarding the structure of syllabics. Similar considerations can therefore apply to the construction of a quality function.

The table T_1 can be replaced with a table of conditional probabilities that if a word λ_i appears in a simple sentence, then the word λ_j will also appear in the same sentence.

"Lexically" similar words should "control" other words according to the same pattern. "Grammatically" similar words, on the other hand, are conversely "controlled" according to the same pattern: e.g., the accusative case is conditioned by the presence of the so-called transitive verbs in the sentence. This is entirely analogous to the starting assumptions used in the construction of the quality function of the algorithm.

This method of analysis is fairly interesting in that it detects "null" morphemes. At a later stage, we will consider a morpheme-identifying algorithm which is unable to identify words. This algorithm, however, will not identify the "null" morphemes, either.

The algorithms probably should be applied in the following sequence: first, we detect the "non-null" morphemes, then words as certain combinations of these "non-null" morphemes, and the words are then again partitioned into morphemes by the above algorithm.

An algorithm for "semantic" classification of words

Classification algorithms are particularly significant for decoding the meaning of a text.

We have mentioned before that external dissimilarity in the appearance of words does not always indicate that the words are significantly different in their meaning. The verbs "to have" and "to possess" are quite dissimilar morphologically, and yet their meaning is evidently very close. It is generally difficult to predict which of the two words, "car" or "automobile," will be used in a sentence, but they are evidently synonymous.

A classification grouping words of similar meaning in an unknown text will greatly enhance the intelligibility of the text. This operation is therefore a necessary step in any decoding routine. As we shall see, a "semantic" classification also provides a quantitative estimate of the "semantic" closeness of words.

We will describe a method proposed by Yu. A. Shreider for developing a system of classifications. In this method, binary and ternary classifications are assumed, i.e., classifications partitioning the entire set of words into two or three classes. For a binary classification, one of the classes contains all the words with a certain common semantic feature, and the other class comprises all the remaining words. In a ternary classification, there is another class of words with an opposite property. Thus, one class of a binary classification may contain words associated with the concept of "space," whereas the other class will contain all other words, unrelated to this concept. The concept of "animation" may constitute the basis of a ternary classification: animate, inanimate, and words to which this feature is inapplicable (e.g., the word "show").

A partial list of useful distinctive features (classifications) is given below: 1) intelligence, 2) elementarity (the property of being unique, 3) action, 4) animation, 5) positiveness (good – bad), 6) greatness (large – small), 7) space, 8) time, 9) order, 10) the property of being a boundary, 11) perception, 12) change, 13) the property of being a part of a whole.

The presence of a certain feature in a given word will be identified by the appropriate number, and the presence of the opposite property will be identified by the number with a bar on top. Some examples of semantic codes of various words are given below:

algorithm 3, 9, 12.
computation process 3, 9, 12.
instant $\bar{6}$, 8, 13.
fool 1, 2, 4, $\bar{5}$.
rest $\bar{3}$, 5, 8, 13.
French plural article 2.

The "semantic codes" can be replaced with vectors, thirteen-dimensional in this case, with the numerals 1, −1, or 0 in the -th position, according as the particular word has the property in question, the opposite property, and no such property:

$$\text{fool } 1, 1, 0, 1, -1, 0, 0, 0, 0, 0, 0, 0, 0.$$

For this thirteen-item semantic classification, closeness is estimated using the equality

$$\rho_K(a, b) = \sum_{i=1}^{n} l(\varepsilon_i, \eta_i),$$

where ε_i is the i-th coordinate of the word a, η_i is the i-th coordinate of the word b,

$$l(\varepsilon_i, \eta_i) = \left\{ \begin{array}{ll} 0, & \text{if } \varepsilon_i = \eta_i, \\ 1, & \text{if } \varepsilon_i = \pm 1, \ \eta_i = \mp 1, \\ 2, & \text{if } \varepsilon_i = 0, \ \eta_i = \pm 1, \\ & \text{or } \varepsilon_i = \pm 1, \ \eta_i = 0, \end{array} \right\} \tag{4.5}$$

Here the distance is chosen so that words with opposite properties are closer to each other than words characterized by presence and absence of a property (e. g., the words "giant" and "dwarf" are closer than the words "giant" and "philosopher"). The distance $\rho_K(a, b)$ is independent of the text: it is determined by the choice of the semantic categories.

A certain textual distance is defined in the following. Using this distance, we can establish a list of categories (the set of classifications).

Let the distance $L(a, b)$ between the words a and b in a sentence be defined as the number of words from a to b, inclusive. (We know that visual distance does not always correlate with distance in meaning. The distance is therefore measured using the so-called graph of the sentence, and not the straightforward text.) Thus $L(a, a) = 0$, and for adjacent words $L(a, b) = 1$. The distance is defined as the mean value of $L(a, b)$ for all sentences with a and b occurring simultaneously. If the words a and b do not occur in a single sentence in a text T, we take $\rho_T^1(a, b) = \infty$.

The function $\rho_T^1(a, b)$ characterizes the distance between "combining" words, whose semantic relation is such that they appear jointly in one sentence. For "mutually exclusive" words, such as "house" and "hut," the function $\rho_T^1(a, b)$ takes on very large values, despite the obvious closeness in meaning of the two words. Mathematically, this shortcoming of the function $\rho_T^1(a, b)$ is manifested in the fact that it does not have one of the basic properties of a metric distance: the "triangle inequality" $(\rho(a, b) + \rho(b, c) \geqslant \rho(a, c))$ is not satisfied for this function. We therefore have to define the distance $\rho_T(a, b)$.

Let

$$\left. \begin{array}{l} \rho_T^2(a, b) = \min\limits_c [(\rho_T^1(a, c) + \rho_T^1(c, b)], \\ \rho_T^k(a, b) = \min\limits_c [\rho_T^{k-1}(a, c) + \rho_T^{k-1}(c, b)]. \end{array} \right\} \tag{4.6}$$

The textual distance $\rho_T(a, b)$ is defined as

$$\rho_T(a, b) = \lim_{k \to \infty} \rho_T^k(a, b).$$

This quantity satisfies all the properties of a normal distance.

Given a certain set of classifications, we can apply $\rho_T(a, b)$ to examine a certain text and hence to improve the original set of categories. This is based on the assumption that the distances $\rho_T(a, b)$ and $\rho_K(a, b)$ are consistent. For example, we may assume that a small $\rho_T(a, b)$ leads to a small $\rho_K(a, b)$, and vice versa.

We will now describe the application of the algorithm. The semantic vectors are defined for a list of words picked out from a given text (Shreider suggests assigning semantic codes to predicate words only, and not to objects, and especially not to proper names, such as "Ivan," "Moscow," etc.). In decoding, all the coordinates of the initial vectors should be zero.

The distances $\rho_T(a, b)$ and $\rho_K(a, b)$ are then determined for these words, and compared with each other. If the consistency criterion is satisfied, the routine is terminated. Otherwise, the system of semantic categories is altered.

Two cases may arise:

1. For some word pairs from $\{a, b\}$, $\rho_T(a, b) < \rho_K(a, b)$. In this case, the coordinates which are different for numerous word pairs from $\{a, b\}$ are eliminated from the list of coordinates (the column vectors). This lowers the dimension of the semantic vectors because the column vectors differentiating between words which are close in a given text are omitted.

2. For some pairs from $\{a, b\}$, $\rho_T(a, b) > \rho_K(a, b)$. In this case, new semantic coordinates are introduced in the following way: we search for a word c such that

$$d = \rho_T(a, c) - \rho_T(c, b) = \max. \tag{4.7}$$

This word c corresponds to a new semantic category (coordinate) which takes the value 0 for the words d satisfying the inequality

$$\rho_T(c, d) < \rho_T(c, b) + \frac{\alpha}{2}, \tag{4.8}$$

and the value 1 for all other words (for the word b this coordinate is 0, and for a, it is 1).

In other words, if the word c is "table," introduction of a new semantic category corresponds to classification of words according to the presence or the absence of the property of "tableness."

We thus obtain a new system of coordinates, different from the original system. It is checked using the consistency criterion and, depending on the results, we pass on to a new coordinate system or terminate the routine.

§10. CLOSENESS-IDENTIFYING ALGORITHMS

Algorithm determining the graph of syntactic connections of words in a sentence

We have mentioned earlier that the visual distance between the elements of a message does not correspond to the intuitive "closeness of meaning." In other words, the intelligibility of a message (i. e., the ability to predict the content of the inaccessible part after examination of the accessible part), although relatively low if the visually observed closeness relations are used, may be markedly enhanced if we pass to optimal relations.

For example, if the message is a linear succession of signals, the successive signals may be regarded as maximally close; however, if the message is a scan of a two-dimensional image, signals separated by a single line length are also maximally close.

Therefore, if the particular signal represents a dark detail in a pattern, a similar black-dot signal can be expected not only in immediate adjacency to the first signal, but also after a certain interval.

Words of an "ordinary" language related in meaning do not necessarily occur one next to the other, either. The arcs in the sentence in Figure 63 [from Walt Whitman's Song of Myself] connect words which are more "closely related" than the unconnected words.

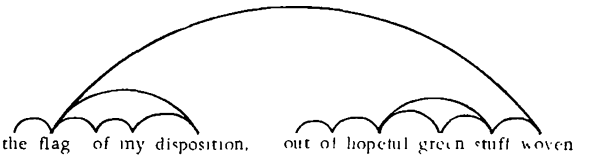


FIGURE 63. Syntactic linkage of words in a sentence "... the flag of my disposition, out of hopeful green stuff woven" (Walt Whitman, Song of Myself).

Pairs of linked words are meaningful, albeit sometimes "half-baked," e. g., "flag of," "of hopeful," "hopeful stuff," whereas other word pairs are meaningless, such as "green woven," "disposition stuff," etc. There is a very close syntactic and semantic linkage between the two extreme words of the sentence, "flag woven."

In some texts (e. g., in Latin poetry), the great visual distance between words related in meaning gives the impression of an intentional jumble (Figure 64).

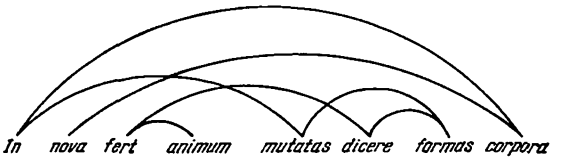


FIGURE 64. Syntactic connections between words in a Latin sentence

Retaining the original word order, the sentence can be translated as follows: "In new attracts the soul changed to tell shapes of bodies," i.e., the soul is drawn to tell how bodies change into new bodies. Knowledge of the true relations is essential for detecting higher level units, since these units consist of "close" units of a lower level.

In describing the basic version of the algorithm below, we shall assume that both the lower-level units — words — and the higher-level units — sentences — are known.

This algorithm should establish a "true" semantic closeness between words in a sentence, or more precisely, in a simple sentence, i.e., a sentence which does not contain other sentences.

Our problem will be solved if we identify pairs of words in a simple sentence which are directly related in meaning.

A pair of words directly related in meaning can be described as a segment of a line whose ends correspond to the particular words or to some symbols replacing the words. Then the entire set of words in a sentence which are directly related in meaning will be represented by a drawing, or what we call a graph, of the general form shown in Figure 65.

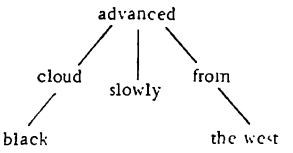


FIGURE 65. Pairs of words connected by straight segments are meaningful.

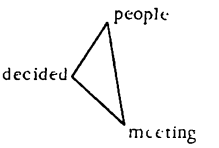


FIGURE 66. A graph with indeterminate meaning.

The shape of a graph characteristic of a simple message can be predicted in advance. The prevailing opinion is that the graph of a simple message is a "tree," or in other words a graph without cycles. A graph is said to be connected if, by retracing its segments (sides) we can pass from any vertex to any other vertex; a connected graph is called a tree if there is only one such path between any two vertices.

This is a reasonable assumption, which indicates that all the words in a sentence are connected in meaning, though possibly not directly. Moreover, these connections are unambiguous, whereas the graph shown in Figure 66 is ambiguous; there are two possible interpretations of the sentence, "decided meeting people" or "people decided meeting."

The basic assumption used in constructing a quality function is that there is a strong predictive link between words which are directly connected in a sentence. However, to apply this principle in practice, we have to set up a certain classification of words. First, words occur fairly infrequently, and pairs of words are even more infrequent. The introduction of classes enables us to group numerous words into a single category, representing the individual words by the appropriate class symbol. The class symbol will occur frequently in the text. Second, the alphabet of words is very extensive. Unless it is compressed, no computer experiments will be able to test the algorithms.

The classes must not be defined at random: "good" classes should be introduced, with the aid of a special decoding algorithm. We cannot propose any particular algorithm of this kind at present, but we have a rough idea of what they should look like.

The algorithm probably will be based on the so-called "grammatic" classes of words. In our specimen calculations for a Russian-language text, we used the following selection of classes: cases of nouns and adjectives, finite verbs, adverbs, verbal adverbs, infinitives, prepositions, conjunctions, particles.

Each word is then assigned the symbol of its class, and the text is decoded using these symbols.

Consider how words (or, more precisely, classes of words) predict one another in a sentence. Let us compute the conditional probability of a word of class k_j occurring in a sentence if it contains a word of class k_i. This conditional probability can be computed from the equality

$$p(k_l/k_j) = \frac{p(k_l, k_j)}{p(k_l)} \approx \frac{P(k_l, k_j)}{P(k_i)} = \frac{\varphi(k_l, k_j)}{\varphi(k_i)}.$$

Here $p(k_i, k_j)$ are the probabilities of the joint occurrence of words of the classes k_i and k_j within a simple sentence, $p(k_i)$ is the probability of occurrence of the symbol k_i in a simple sentence. P and φ denote relative and absolute frequencies, respectively.

To estimate the "mutual predictability" of the symbols k_i and k_j, we form the average of $p(k_i/k_j)$ and $p(k_j/k_i)$, denoting it $\bar{p}(k_l, k_j)$.

A partial predicate system for a given text is defined as the square table of numbers $\bar{p}(k_i, k_j)$.

If a certain sentence tree is given, we can assign a weight $p(k_i(\lambda_u), k_j(\lambda_v))$ to every side of the graph joining the words λ_u and λ_v. This weight is a function of the classes k_i and k_j and the words λ_u and λ_v at the two endpoints of the segment.

It is assumed that words directly connected in a graph are characterized by a high mutual predictability. We may therefore use as the quality function of the graph the sum of the numbers $\bar{p}(k_i(\lambda_u), k_j(\lambda_v))$ for all the sides of the graph.

The set of permissible solutions in our case is thus the set of all the possible trees constructed from the symbols of the given simple sentence, and the quality function is

$$D = \sum_u \sum_v \bar{p}(k_l(\lambda_u), k_j(\lambda_v)),$$

where λ_u, λ_v are the symbols joined by the particular side of the tree. The best solution maximizes the function D.

A similar problem has been tackled in mathematical economics to find the shortest interurban telephone network. There are two versions of the algorithms extremizing the function D. We will describe here the simpler of the two, published in /6/. This method, like most other algorithms of the so-called discrete analysis, is not exact, but it is quite acceptable in practice. In principle, a rigorous and exact algorithm can be devised, which would examine all the possible trees, calculate the corresponding values of D, and choose the tree which maximizes D. For long sentences, however, this method is too cumbersome to be practicable.

The set of word symbols of a given sentence is partitioned into two parts, accessible and inaccessible. Initially, the accessible set is empty.

The first step is to choose two vertices λ_1 and λ_2 such that

$$\bar{p}(k(\lambda_1), k(\lambda_2)) \geqslant \bar{p}(k(\lambda_i), k(\lambda_j)),$$

where λ_i and λ_j are two words of the text which do not coincide with λ_1 and λ_2, respectively. The vertices λ_1 and λ_2 are accessible and are joined by a side of the graph.

If there are both accessible and inaccessible vertices, we search for an inaccessible vertex λ_{in} such that

$$\bar{p}(k(\lambda_u), k(\lambda_{in})) \geqslant \bar{p}(k(\lambda_v), k(\lambda_w)),$$

where λ_u is some accessible vertex, λ_w is any inaccessible vertex other than λ_{in}, and λ_v is any accessible vertex. The vertices are joined by a side of the graph, and λ_{in} is added to the list of accessible vertices. If there are no inaccessible vertices, the routine is terminated.

This algorithm has a number of more interesting versions. We can indicate the order of best reading of the sentence, i.e., the examination procedure which ensures the fastest decoding of meaning. The "main ideas," or more precisely, the main words are the first to be grasped in this way. We can thus fix a certain order of preference or a subordination relation for the words in the sentence.

To this end, it suffices to indicate directions on the sides of the graph connecting the different words. These directions are determined from the following considerations: a subordinate word more strongly predicts its principal, is "in a greater need of the principal," than the other way round. Therefore, if $p(k(\lambda_i)/k(\lambda_j)) > p(k(\lambda_j)/k(\lambda_i))$ the arrow should point from λ_j to λ_i.

If we read the sentence in the reverse direction, the sense of unintelligibility and confusion will persist for a longer time, since the already examined subordinate words strongly predict their principals, which still remain inaccessible.

Anyone who has ever tried to learn German and Latin is familiar with this curious feeling!

An algorithm identifying "types of syntactic relationship" of words

The algorithm described above has been checked manually for small examples only. However, before proceeding with serious experiments, we should analyze some of the errors which this algorithm introduces.

Apparently, most of the errors are associated with the extreme imprecision of the description of the different words: after all, even words of the same grammatic class are markedly different from one another. For example, the sentence "structure of hydrogen atom" [in Russian – stroenie atoma vodoroda] is coded as n_n, g_n, g_n.* If

* Nominative case of a noun, genitive case of a noun, genitive case of a noun.

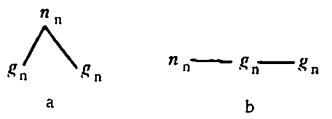


FIGURE 67. Incorrect (a) and correct (b) graphs of the Russian sentence "structure of hydrogen atom" — "stroenie atoma vodoroda."

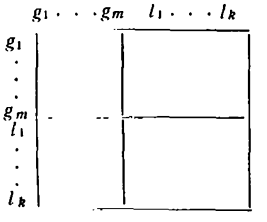
$\bar{p}(n_n, g_n) > \bar{p}(g_n, g_n)$ (as is to be expected), the algorithm will produce the tree shown in Figure 67a, whereas the correct graph of this sentence is b in the same figure.

It is readily seen that connectedness of words is sometimes independent of their grammatic form; for example, the graph for the part of a sentence which reads "by structure of hydrogen atom" [stroeniem atoma vodoroda] in Russian is evidently independent of the case of the first word and therefore coincides with the graph of the original sentence "structure of hydrogen atom" [stroenie atoma vodoroda].

The algorithm described below is intended for avoiding these errors; it is also of independent interest.

Suppose that words are described using two classifications, and not one as before. One of these classifications is a grammatic classification, and it will be identified with the classification previously described. Another classification is a lexic classification. We assume that the classes of the lexic classification contain words with a common base. Each word is thus described by two symbols: the symbol of its grammatic class and the symbol of the lexic class.

A "partial predictive system" for this description of a dictionary can be presented in the form of a table where the row g_i contains numbers characterizing the grammatic class g_i, and the column under l_j contains numbers characterizing the lexic class l_j. The entry in the square i, j in the table is the conditional probability that a sentence containing a word of class i will also contain a word of class j. The numerical entries are calculated by analyzing a certain text



Once the table has been computed, we can calculate the so-called connecting function σ_1 for any two words λ_i and λ_j:

$$\sigma_1(\lambda_i, \lambda_j) = p(g(\lambda_j)/g(\lambda_i)) + p(g(\lambda_j)/l(\lambda_i)) + p(l(\lambda_j)/g(\lambda_i)) + p(l(\lambda_j)/l(\lambda_i)).$$

This function coincides with the conditional probability, but it has the advantage of enabling us to described words in terms of two different classifications.

We have already noted that the lexic grammatic classes coincide with the classifications of syllabics into vocalic and consonantal. A given pair of classifications therefore can be identified in principle by an algorithm similar to that mentioned on p. 191.

As in the first version of the algorithm, the table of conditional probabilities is replaced by the table of mutual predictabilities, i. e., numbers of the form $\bar{p}(k_i, k_j) = \frac{1}{2} p(k_i/k_j) + p(k_j/k_i)$; the connection function σ_2 calculated using the table $\|\bar{p}(k_i, k_j)\|$ takes the form

$$\sigma_2(\lambda_i, \lambda_j) = \bar{p}(g(\lambda_i), g(\lambda_j)) + \bar{p}(g(\lambda_i), l(\lambda_j)) + \\ + \bar{p}(l(\lambda_i), g(\lambda_j) + \bar{p}(l(\lambda_i) l(\lambda_j)).$$

Consider a sentence where every word has been coded with a pair of symbols of this kind. We can then construct a tree of this sentence assigning the numbers $\sigma_2(\lambda_i, \lambda_j)$ to the sides of the graph and applying the procedure described above.

Errors of the kind encountered in the sentence "structure of hydrogen atom" are corrected because the connecting function allows for the frequent joint occurrence of the two words "hydrogen atom" in a single sentence, irrespective of the cases (this fact is expressed by the high value of the term $\bar{p}(g(\lambda_j)/l(\lambda_i))$. The sides of the graph are directed using the same considerations as before.

This algorithm yields qualitatively new information about sentences. We can establish four types of connection, depending on which of the terms makes the largest contribution to the value of the connecting function σ_2, $p(g(\lambda_j)/g(\lambda_i))$, $p(g(\lambda_j)/l(\lambda_i))$, $p(l(\lambda_j)/l(\lambda_i))$ or $p(l(\lambda_j)/g(\lambda_i))$. Their interpretation is readily understood if we remember the concept of "subordinate clauses" taught in high school. Knowledge of the various connections enhances the intelligibility of the text, since it provides the possibility of direct prediction of words. Let us consider the different types of connections.

I. $g \rightarrow g$ type. A form of connection whereby the grammatic class of a word λ_i strongly predicts the grammatic class of a word λ_j, an effect generally called consistency.

II. $l \rightarrow g$ type. The lexic class of a word λ_i strongly predicts the grammatic class of a word λ_j. In traditional grammar this is known as "government."

III. $l \rightarrow l$ type. The lexic class strongly predicts the lexic class of another word. The adjoining effect of traditional grammar. According to the traditional approach, this type of connection is characteristic of words which do not change in declension, e. g., "very early" (it is not clear to what extent this definition is consistent with the traditional definition of the concept; the same holds true for the other types, however).

IV. $g \rightarrow l$ type. Without analogy in traditional grammar. May not occur in reality altogether. No obvious examples known.

The simplest algorithm of literal machine translation

We will now consider one last generalization of the sentence-graph algorithm, which although of unquestionable importance, requires a great deal of preliminary information about the text.

We will show that this version of the algorithm may be useful for machine translation (possibly after some modifications). We mean here the simplest and most elementary form of machine translation, the so-called literal translation, when the translation process assigns a certain word to every significant word of the original.

The difficulty in this translation is that the same word of the original lends itself to different translations. If there is a dictionary which assigns to the words of the source language all the possible translations, the main problem of the translation algorithm is to reject all the redundant alternatives, retaining only the most accurate and fitting translation. The higher the number of the alternatives rejected, the more sophisticated is the algorithm.

The complete dictionary should translate the English word "hand" into Russian as "кисть" (human hand), "стрелка" (of a clock), and also give the various declensions, etc. In this case, the choice of the best alternative would involve choosing the correct word in a correct grammatic form. This routine requires a system of semantic classifications, possibly similar to that described on p. 192, and also certain "grammatic classifications." We should be able to describe the words of the translation language using this classification system, i. e., to every word of the translation we should be able to assign its description vector (see p. 193). The list of symbols of the semantic classes will constitute a certain "semantic" alphabet.

For the translation language we then can construct a square table whose rows are identified with the symbols of classes and the entries are the conditional probabilities $p(k_j/k_i)$ (this table is computed using an extensive text in the translation language). The connecting functions can be defined in the same way as before:

$$\sigma_3(\lambda_i, \lambda_j) = \sum_{u=1}^{n} \sum_{v=1}^{n} p\,(k_u(\lambda_j)/k_v(\lambda_j)),$$

where n is the total number of classes:

$$\sigma_4(\lambda_i, \lambda_j) = \sum_{u=1}^{n} \sum_{v=1}^{n} \bar{p}\,(k_u(\lambda_i), k_v(\lambda_j)).$$

The function σ_4 takes on large values for pairs of words which strongly predict each other in the translation language.

To describe the remaining steps of the algorithm, consider a schematic diagram of some sentence in the source language and the alternative translations performed with a bilingual dictionary.

	First word	Second word	Third word
Translation alternatives	+		+
	+		
	+	+	+
	+		

One word should be picked out from each column. If every word of each column is joined by a line with every word of all other columns, we obtain a graph which contains all the possible graphs of all the possible sentences (Figure 68).

From this graph, however, we should select only one tree, with one vertex in each column. This graph can be found in the usual way (with slight modifications); the only difference is that we are not looking for a graph connecting all the vertices: in each column, all the vertices but

one should be isolated, as in Figure 69. The isolated vertices are the rejected alternatives.

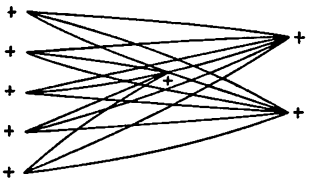


FIGURE 68. A graph containing all the possible graphs of the translation sentence.

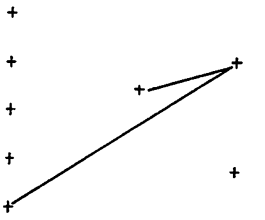


FIGURE 69. A probable end product of a graph-selection algorithm.

§11. MATCHING ALGORITHMS (END)

Morpheme-identifying algorithms

As we have noted before, morphemes are the smallest meaningful units of the human language, i. e., the set of morphemes is the alphabet of elementary signals for semantic levels.

The word "unloader," for example, is divided into morphemes as "un+ load+ er." Certain morpheme sequences form words. Words in various orthographies, although by no means in all, are separated by special signs — blanks — from one another. Morphemes, as a rule, are not separated from one another. Therefore, in automatic text analysis, the word boundaries are assumed to be known, whereas the boundaries between morphemes are sought by a special algorithm.

However, we have no right to expect a message from space to follow the pattern of terrestrial writing. We shall therefore have to aim at the most difficult case, namely when neither the word boundaries nor morpheme boundaries are defined to start with. In such a case, we should start with an identification of code groups. The main difficulty is that the code groups forming morphemes are of varying length. Therefore, the set of permissible solution is the set of all possible partitions of the text, the number of which is $(N-1)!$ There is one further significant difference: the number of morphemes in human languages is much greater than the number of letters which represent sounds. This is not a fundamental difference, but a simpler procedure would probably apply to the decoding of an inhomogeneous code representing letters.

The algorithm described in what follows uses a quality function which is different from V and a different recognition procedure. Because of these differences, the algorithm will identify higher level groups of letters, and not only morphemes. Groups of morphemes are of interest if they

constitute words or groups of words. The algorithm described below identifies not only morphemes, but also words (at least some) and certain groups of words.

There is of course a possibility that certain combinations of morphemes will remain "unresolved"; a special algorithm will be required to partition these combinations into the constituent morphemes.

Because of the change in the purpose of the algorithm, the set of permissible solutions is different. A correct partition of the text by round parentheses is defined as the original text with symbols of two kinds — right and left parentheses — interposed. The right and the left parentheses are placed in accordance with the following rules:

1) a correct group of zero order is a group of letters of the original text enclosed in a left parenthesis on the left and in a right parenthesis on the right;

2) a correct group of i-th order is a group of correct groups of $(i-1)$-the order.

A correct partition of a text by round parentheses is obtained if the parentheses interposed in the original text convert it into a regular group.

It follows from this definition that correct groups of the same order may "link up," i. e., the beginning of one regular group may coincide with the end of another group, if the one is not contained completely in the other. This concept establishes the complete separability of morphemes, which are never intertwined into a jumble in a sentence.

Simple groups correspond to morphemes; this, however, does not take into consideration the possible inclusion of morphemes in one another (man — men) and other exceptional effects. It is impossible to allow for these effects without markedly complicating the algorithm.

The quality function will be derived from the following considerations. While analyzing the previous algorithm, we noted that for code groups of equal lengths, the difference between the groups of correct and incorrect solutions is that the parts of correct groups are "connected" more closely than the parts of incorrect groups. This concept constitutes the basis of the new quality function.

A group is said to consist of "strongly connected" parts if the appearance of a certain part of the group strongly predicts the appearance of the remaining part. For example, if a certain text contains the group "envelo," then it is almost certain to be followed by "pe." This proposition holds with varying degree S of likelihood for other partitions of the word "envelope." The predictability is naturally calculated using the expression for the conditional probability: given an i-th partition of a word λ into right and left parts, we can write the word in the form $\lambda = L_i R_i$ (where L_i is the left part of the i-th partition, R_i is the right part); the predictability of the right part by the left part of the word can be expressed by the fraction $\frac{\varphi(L_i R_i)}{\varphi(L_i)}$. The predictability of the left part by the right part is similarly expressed by the fraction $\frac{\varphi(L_i R_i)}{\varphi(R_i)}$.

On the whole, the quality of a group is determined as follows:

1. The group is divided into two parts in all possible manners.
2. For each partitition, the predictability of one of the parts by the other is determined.

3. The mean of all the predictabilities corresponding to the various partitions is calculated.

Since in a chain of length d we may introduce $d-1$ partitions, the total number of predictabilities (i. e., fractions of the form $\frac{\varphi(L_iR_i)}{\varphi(L_i)}$ and $\frac{\varphi(L_iR_i)}{\varphi(R_i)}$) is $d-1$. The quality of a group or its "stability" is therefore expressed by the equation

$$Y_\lambda = \frac{1}{2(d-1)} \sum_{i=1}^{d-1} \left(\frac{\varphi(L_iR_i)}{\varphi(L_i)} + \frac{\varphi(L_iR_i)}{\varphi(R_i)} \right).$$

This relation is naturally inapplicable to evaluating single-letter groups: the stability of these groups is taken equal to zero. The stability of groups which do not occur in the text or which occur only once is also set equal to zero; the latter choice is explained by the fact that any group of the text including an incorrect one, occurs at least once.

The quality of the partition $Y(R)$ is calculated as the sum of stabilities of all the correct groups entering the particular partition.

It would naturally be impossible for us to examine all the correct partitions of a text by round parentheses and to calculate for each partition the quality function $Y(R)$. We will therefore propose a routine which, as always, ensures a fairly high value of $Y(R)$.

We will require a frequency dictionary of all the groups occurring in the text, i. e., a list of these groups indicating the number of occurrences of each group in the text.

The simplest method for compiling such a dictionary is the following: first we draw up lists of all the possible groups of length 1, 2, . . . , etc., and then examine the text and count the number of times the particular group occurs. This method, however, is unsuitable for fairly long groups.

The compilation of the frequency dictionary is substantially simplified since we are not concerned about groups which occur less than twice (their stabilities are zero). We can therefore use the following routine: draw up a list of letters and determine their frequencies; letters occurring less than twice are dropped from the list; the frequencies of the remaining letters are included in the final list. The resulting fragment of a frequency dictionary is called a first-order fragment.

If a fragment of order $i-1$ has been compiled, a fragment of order i is constructed in the following way: for every group λ from the fragment of order $i-1$, we construct all the possible groups of the form λa_x where $a_x - 1$ is some letter of the alphabet; then determine the frequencies of these groups and omit all chains which occurred less than twice. The list of remaining groups and their frequencies constitutes a fragment of order i.

The successive construction of fragments is terminated when a fragment of the next higher order proves to be empty. This stage is reached when all the groups longer than any of the previously constructed groups occur less than twice in the text.

The frequency dictionary can be markedly shortened by omitting all the groups which are contained in some longer group of the same frequency. If some group λ_u occurs k times, any group λ_v contained in λ_u occurs at least k times; therefore if the frequencies of λ_u and λ_v are equal, there is no need to include the group λ_v in the dictionary.

Once sufficient information is available on group frequencies, we can find their stabilities. The computed stabilities are also included in the frequency dictionary together with the corresponding frequencies.

The frequency dictionary of groups is then converted into a frequency dictionary of correct groups.

The assumption which ensures the first step of the routine amounts to the following: it is assumed that the most stable group always enters the text correctly, i. e., it does not link up with any other correct group.

To explain this assumption, note that some groups form morphemes or combinations of morphemes in certain parts of the text, and not in others. For example, "un-" is a morpheme in "unloader," but it is not a morpheme in "bunloader."

In the frequency dictionary, we should thus find the most stable group λ_1, and then enclose in parentheses each occurrence of this group in the text.

For the same reason, none of the groups linked up with any of the inclusions of the group λ_1 may be a correct group. We should therefore reduce the frequencies of all the groups in the dictionary whose inclusions are linked up with the inclusions of λ_1; the reduction should be equal to the number of linkages of the corresponding groups with λ_1.

Groups whose frequencies should be reduced can be found by considering the inclusions of λ_1 according to the following scheme:

$$a_i (a_{i+1} \dots a_{i+l-2} a_{i+l-1} \overset{k}{[} \dots \overset{21}{[[} a_{i+l}) a_{i+l+1}] \overset{1}{} a_{i+l+2}] \overset{2}{} \dots \overset{k}{]} \dots$$

Here the group $a_{i+1} \dots a_{i+l}$ represents the inclusion of a correct group in the text; i is the running number from the beginning of the text.

In the frequency dictionary, we locate the groups enclosed in square brackets in the order of their increasing numbers; $k + 1$ is the number of the first group which has not been entered into the frequency dictionary.

The frequencies of those groups which are located in the dictionary are reduced by 1; the square brackets are then extended to the right of the interval between a_{i+l-1} and a_{i+l-2}, and then to the right of the interval between a_{i+l-2} and a_{i+l-3}, and so on, until we reach the interval between a_{i+1} and a_{i+2}. Similar series of brackets are interposed to the left of the group $a_{i+1} \dots a_{i+l}$.

All the inclusions of the group λ_1 are examined in this way. As a result of the application of this procedure, the frequency dictionary approaches a list of correct groups, since the number of interlinking groups in the text diminishes. If the reduction in frequencies brings the frequency of some group below 2, the particular group is omitted from the dictionary.

In general, a frequency dictionary can be considered as an approximation to a list of regular groups of finite stability. The quality of the list can be estimated using the relation

$$Y' = \sum_i Y(\lambda_i) \varphi_{cor}(\lambda_i).$$

Here λ_i is a certain group, $\varphi_{cor}(\lambda_i)$ is the number of correct inclusions of λ_i in the text.

values of the original function, $\tilde{x}(t_k) = x(t_k)$. The question is, are these functions equal at any time t, and not only at t_k, i.e., are they identically equal? The fit between the two functions is naturally improved if the original function varies slowly between the quantization times t_k. This means that the function should not contain very high harmonics. According to Kotel'nikov's theorem, the two functions are identical if the original function $x(t)$ does not contain components with frequencies ν higher than Δf, i.e., if the band width of the $\Delta\nu$ transmitted function is equal to the band width of the communication channel. Kotel'nikov's theorem is highly significant for the theory and technology of communication, since it permits converting continuous functions into a train of some discrete magnitudes for transmission. This theory maintains that a function with a bounded spectrum $\Delta\nu$ is completely determined by its values measured at intervals $\Delta t = {}^1/_2\Delta\nu$. In particular, a function of duration Δt, i.e., a function which does not vanish only for $t_0 < t < t_0 + \Delta t$, is determined by a set of $2\Delta t\Delta f$ discrete values. Thus, the definition of information derived for discrete messages can be safely applied to continuous functions with a bounded spectrum.

When continuous functions are transmitted by means of pulsed signals, the main difficulty is that the function may take on any instantaneous values, including irrational and transcendental numbers with an infinite number of significant digits. Theoretically (in a noise-free channel), these numbers can be transmitted with full faithfulness by PAM or another suitable technique. In reality, however, reconstitution of the original pulse with sufficient accuracy (or transmission of a sufficiently high number of significant digits) in a noisy channel requires an excessively high signal-to-noise ratio in the communication channel. Therefore, the next step adopted in the transmission of continuous functions calls for quantization of the message. To quantize the message, we select from among all the values of $x(t)$ a set of N discrete allowed levels $x_1, x_2, \ldots, x_N$, which are distant Δx from one another (the quantization gap). All the other values are regarded as forbidden. Only the allowed values are transmitted. If the true instantaneous value of the function falls inside the interval (x_i, x_{i+1}), i.e., takes on a forbidden value, the nearest allowed value, differing from the true value by less than half the quantization gap, is transmitted through the channel. This operation is completely analogous to the rounding-off of numbers; it essentially signifies that we are transmitting the true values of the function up to a certain number of significant digits.

The quantized values of the signal in the communication channel are affected by random noise. The width of the quantization gap should be so chosen that with a given probability p the noise does not exceed half the quantization gap. Then the signal can be accurately reconstituted at the receiving end of the channel, since in this case the signal level nearest to the noise-distorted value is the same as that fed into the communication channel. The probability of signal reconstitution error is equal to the given value p. The reconstituted signal can be again sent through the communication line, and this procedure may be repeated several times, without affecting

Letter-identifying algorithm

We mentioned on p. 148 that linguistic phenomena are best characterized by a selection of distinctive features, but the question of how to choose the set of distinctive features still remains open. Consider the following situation: there exists a set of "elementary features." These elementary features are then combined in an optimum manner according to certain rules to produce compound features. A certain criterion (e. g., a quality function) is required to assess the quality of these features.

The simplest part of this program is apparently the choice of the set of elementary features. The particular choice, it seems, would be largely immaterial.

Let us consider the elementary features for messages received visually. This category includes written texts, pictures, and the entire visible universe. It is clear that the elementary features should constitute the simplest elements of sensation associated with the smallest differences noticeable to the eye (detectable with our detector).

If every color is regarded as a mixture of the three basic colors — red, blue, and yellow — the visible universe can be represented as a six-dimensional space, with each point described by the six coordinates $\alpha_r \cdot \varepsilon_r$, $\alpha_b \cdot \varepsilon_b$, $\alpha_y \cdot \varepsilon_y$, $\alpha_l \cdot \varepsilon_l$, $\alpha_w \cdot \varepsilon_w$, $\alpha_h \cdot \varepsilon_h$ where ε_r, ε_b, ε_y are the minimum increments in the intensities of the three colors, ε_w, ε_l, ε_h are the minimum detectable widths, lengths, and heights, expressed, say, in angular units. The symbols containing the letter ε are unit vectors or "elementary features"; symbols containing the letter α correspond to the strength or the magnitude of the feature in the particular object.

Similarly, acoustic impressions can be described in terms of elementary features associated with minimum detectable acoustic differences, but this approach is not absolutely essential: sounds can bedecoded and presented in chart form (e.g., an oscillogram).

In general, we receive information about the outside world in the form of what is known as "sensation"; signals which are not detected directly by our sensory organs are first converted by "physical instruments."

On the other hand, the transition from microevents (points in the space of elementary features) to more complex units, e. g., letters in the usual sense, is in general a highly complex problem.

We will now consider, in a highly approximate form, the problem of letter identification, assuming that we have in our possession an instrument which is capable of distinguishing between black and white squares (on a sheet of paper covered with a fine grid) and identifying the position of the squares.

This problem belongs to the domain of pattern recognition. Extensive literature is currently available on the subject.

Note, however, that the great majority of sources treat the problem from the aspect of learning to recognize objects. Procedures based on this principle are constructed as follows: consider a certain set of objects presented to the computer; the computer has a certain number of responses. The computer should assign one of its responses to every object (i. e., in practice, it should generate a certain classification of the objects). If the machine "errs," the teacher — a human operator — informs the computer

of its error and "penalizes" it; otherwise, the computer is "rewarded." When the learning stage is completed, the computer is capable of recognizing the objects independently.

The learning approach is naturally ruled out in decoding problems. Moreover, simple classification of objects is not enough in our problems: the boundaries between different letters must be indicated.

If we ignore for the moment the problem of combining the "images" of the same letter into classes, the problem of letter identification becomes similar to the problem of identification of simple morphemes. We may assume with fair accuracy that letters occupying non-overlapping parts of the surface and the combinations of dots filling these areas are stable in a certain sense.

Letters are not necessarily similar to meaningful images. Letters need not be smooth or connected: they may represent a random combination of points. It suffices that all the inclusions of one letter correspond to roughly the same combination of points.

A distinctive feature of the problem of letter identification compared to the problem of morpheme identification is that no two identical inclusions of a single letter exist: there are only similar inclusions of varying likeness. The basis for this likeness is extremely difficult to detect: sometimes a slight change in the outline coverts one letter into a different letter (e.g., the Russian letters И and Й, Ш and Щ), whereas much more radical morphological changes leave the letter unchanged (σ and ς, ∂ and d). The size of letters, their inclination, and the degree of stretching generally do not matter, although the cursive e and l differ in size only, the Cyrillic Е and Ш are in fact the same pattern rotated clockwise through 90°, p and b have the same shape, rotated in a plane and reflected.

In certain writing systems, e.g., shorthand, even less conspicuous features are decisive for letter pronunciation, e.g., elevation above the line level, thickness of strokes, etc.

The most effective algorithm should apparently contain a set of rules for building up elementary features into really distinctive features of letters. Some of these features are probably recognizable only in the presence of other letters used for comparison (e.g., elevation above line level, inclination, size).

Letter identification without reference to nearby letters is therefore an impossible undertaking. We can only hope that in most cases it suffices to examine a very small neighborhood of the letter being analyzed.

The algorithm described below does not pursue any serious aims. Nevertheless, it is of a certain interest because it uses a very limited amount of initial information (in particular, letter sizes need not be considered) and leads to a definition of frequency which is far from self-evident.

Let us first define the so-called residual similarity. Let an element a be located in some area K_1 of the text. Suppose that this element is described by a function $\delta(x, y)$, where x and y are the rectangular coordinates of the points in that area, and $\delta(x, y)$ takes on two values only: 1 for a black dot, and 0 for a white dot.

Let an element b be located in another area K_2 and suppose that this element is described by a function $\delta_1(x_1, y_1)$, where x_1, y_1 are the coordinates of the points of the second area in its own system of axes. These are also rectangular coordinates.

The coordinate axes of the two areas are translated until their origins coincide and the respective axes are parallel. The residual dissimilarity $\frac{1}{S}$ is defined by the expression

$$\frac{1}{S} = 1 + \frac{1}{|K_1 \cap K_2|} \iint_{(K_1 \cap K_2)} (\delta_1(x_1, y_1) - \delta(x, y))^2 \, dx \, dy.$$

The symbol S stands for "residual similarity." The symbol $K_1 \cap K_2$ indicates intersection of the corresponding areas.

The following transformations are allowed for the coordinates of the area K_2: 1) parallel translation, 2) similarity transformation, 3) contraction along each of the axes, 4) change of angle between the axes, 5) rotation in a plane, 6) mirror reflection.

The function δ_1 and the coordinates x_1, y_1 are transformed so that the values of the new function in the new coordinates coincide with the values of the old function in the old coordinates, i. e.,

$$\delta_1'(x_1', y_1') = \delta_1(x_1, y_1),$$

where $\delta_1'(x_1', y_1')$ stands for the new function in the new coordinates. Under these conditions, the change in the function will be completely determined if we specify the transformation of coordinates. If a new function is expressed in the old coordinates and the residual similarity with $\delta(x, y)$ is determined, we will find that it has changed compared to the residual similarity before the transformation. Thus, each transformation of the original coordinates can be assigned a certain value of the residual similarity.

Let the six transformations define six axes in the transformation space. Along each axis we lay off the "values" that each transformation may take (it is assumed that a single-valued quality function has been defined for each transformation). The points of this space, defined by combinations of the values of the quality functions, will be called permissible points. At every permissible point, two functions are defined: a scalar function — the residual similarity, and a vector function — the gradient indicating the direction of fastest growth of the residual similarity. Moving along the gradient, we can find a point at which the residual similarity reaches its maximum value for the two areas and the given elements. This maximum value will be used as the true similarity of the two elements a and b. We will use the same symbol S as before, and in the following S is to be interpreted in this sense.

Consider an arbitrary area K of the text, to be used as a reference standard. The contour enclosing this area is translated by the smallest possible steps in the vertical and horizontal direction. For every position of the contour, we determine the similarity of the corresponding area to the original area. S in this case is a function of the position of the contour. The number of maxima of this function is adopted as the frequency of the element contained in the initial text area. Consider a text area made up of two squares K_1 and K_2, with a common side. The absolute frequencies of the first and the second square and the frequency of the rectangle made up of the two squares will be designated $\varphi(K_1)$, $\varphi(K_2)$, and $\varphi(K_1, K_2)$.

The predictability of K_2 from K_1, or $p(K_2/K_1)$, is defined as the ratio $\frac{\varphi(K_1K_2)}{\varphi(K_1)}$, and the predictability of K_1 from K_2 is correspondingly defined as $\frac{\varphi(K_1K_2)}{\varphi(K_2)}$, or $p(K_1/K_2)$. The average of these two predictabilities is defined as the mutual predictability of K_1 and K_2, or $p(K_1, K_2)$.

We now start increasing the size of K_1 and K_2, measuring the predictability after every small increase (the straight line accommodating the boundary and the position of the center of the boundary remain unchanged). We thus present the mutual predictability as a function of the size of the squares. A plot of this function is shown in Figure 70, where D is the size of the rectangle.

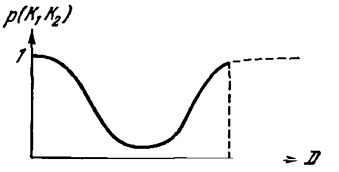


FIGURE 70. A probable plot of the function $\overline{p}$ (K_1, K_2) vs. the size of the squares K_1 and K_2.

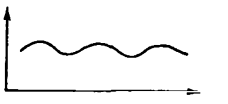


FIGURE 71. A probable plot of the function C vs. boundary position.

Indeed, all the mutual predictabilities are less than unity: after all, the symbols contained in pairs of squares are more individualized and richer in detail than symbols enclosed in separate squares. Elements showing maximum to symbols enclosed in separate squares are therefore more frequent than symbols in pairs of squares.

If the squares are small, the mutual predictabilities are close to unity, since the elements in separate squares and in pairs of squares are still similar to one another, and their similarity functions have maximum at approximately the same points of the text.

As the size of the squares increases, the mutual predictabilities at first decrease, and then increase reaching unity, since the elements in large separate squares, and likewise the elements in a pair of squares (a rectangle) occur once only.

Let us find the minimum value b of the variable D such that for any choice of the boundary between the squares K_1 and K_2 in the text (with squares of the size b), the mutual predictability $p(K_1, K_2)$ is unity. The centrality of the boundary of the squares K_1 and K_2 is defined as $C = \int_0^b p(K_1, K_2)(D)\,dD$. The centrality characterizes the mutual predictability of the squares, irrespective of their size, and is thus a fundamental characteristic of the boundary location.

The basic hypothesis of our algorithm is the assumption that centrality is minimized (compared to other near positions of the boundary) when the boundary passes between the letters and is aligned along the true "physical" boundary of the letters. Then we start moving the boundary in some direction, measuring the values of C after each step. We obtain a plot of the form shown in Figure 71.

Fixing the boundary at a point corresponding to a local minimum, we will rotate it about its midpoint, measuring C at minimum angular intervals. We select the angle α corresponding to a minimum value of C and move the boundary infinitesimally in the given direction. In the new position, we again select the best direction for further displacement, and act as before. If several best directions are available, we choose the rightmost. If we are really moving along the letter contour, we will describe a closed curve enclosing the letter completely.

The element enclosed within the curve may be used as a reference standard, and we then calculate all the maxima of S in the text using this standard.

The textual elements identified by this procedure are not quite the letters of the alphabet. For example, the letters p and b are definitely different, and we will therefore call this a skeleton alphabet.

Without going into details, we can outline a general procedure for developing this skeleton into a proper alphabet. Each occurrence of a skeletal element is assigned a vector from the transformation space which transforms it, say, into the first occurrence of the element in the text. Some of the component values of this vector are stable, i.e., strongly predictable by the nearby elements. For instance, the letters of the word "bed" strongly predict the variety of the symbol "b" with the stroke directed upward, whereas the letters of the word "pet" predict the same symbol with a downward stroke. Accidental changes of symbols are not predicted with any stability.

§12. CONCLUSION

In conclusion, we wish to call the attention of the reader to one remarkable aspect of decoding.

How strange to our conception may the reality hidden in extraterrestrial messages be? Will we at all be in a position to comprehend the content of these messages?

It should be stressed at the outset that there is an enormous difference between understanding the message and comprehending what it is all about. While intelligibility is based on the ability to predict the inaccessible parts of the message or future events, comprehension draws upon our ability to translate the message into the language of images corresponding to real situations.

Not all that is intelligible is comprehensible. We cannot comprehend the sensations of a being which responds to radio frequencies, but this does not detract from the intelligibility of his behavior. Therefore, even if the "other" reality is fantastically strange in our eyes, it need not be considered unintelligible.

Bibliography

1. Apresyan, Yu. D. Idei i metody sovremennoi strukturnoi lingvistiki (Methods and Ideas of Modern Structural Linguistics). Moskva. 1966.
2. Hrozný, F. Khettskie narody i yazyki (Hittite People and Languages). — Vestnik Drevnei Istorii, Vol. 2, No. 3. 1938.
3. Dobrushin, R. L. Matematicheskie metody v lingvistike (Mathematical Methods of Linguistics). — In: "Matematicheskoe Prosveshchenie," Vol. 6, Moskva. 1961.
4. Kaplan, S. A. Elementarnaya radioastronomiya (Elements of Radio Astronomy). — "Nauka." 1966.
5. Pratt, F. Secret and Urgent. — N. G. 1942.
6. Prim, R. K. Shortest Communicating Networks. — Kiberneticheskii sbornik, Vol. 2. [Russian translation. 1961.]
7. Romanov, V. P. Integral'nye metody opoznavaniya. Chitayushchie ustroistva (Integral Recognition Methods. Reading Devices). — Moskva. 1962.
8. Sukhotin, B. V. Algoritmy lingvisticheskoi deshifrovki (Algorithms of Linguistic Decoding). — Problemy strukturnoi lingvistiki. Moskva. 1963.
9. Sukhotin, B. V. Eksperimental'noe vydelenie klassov bukv s pomoshch'yu elektronnoi vychislitel'noi mashiny (Experimental Identification of Groups of Letters with a Computer). — Problemy strukturnoi lingvistiki. Moskva. 1962.
10. Sukhotin, B. V. Algoritm sravnivayushchii bukvy dvukh razlichnykh yazykov (Algorithm Comparing Letters of Two Different Languages). — Nauchno Tekhnicheskaya Informatsiya, Vol. 2. 1966.
11. Sukhotin, B. V. Issledovanie yazyka deshifrovochnymi metodami (Language Studies by Decoding Techniques). — Russkii Yazyk v Shkole, Vol. 6. 1966.
12. Harris, Z. From Phoneme to Morpheme. — Language, Vol. 28, No. 1. 1952.
13. Halle, M. Fonologicheskaya sistema russkogo yazyka (Phonological System of the Russian Language). — Novoe v lingvistike. 1962.
14. Shannon, C. E. and W. Weaver. The Mathematical Theory of Communication. — Urbana. Univ. of Illinois Press. 1959.
15. Shreider, Yu. A. Mashinnyi perevod na osnove smyslovogo kodirovaniya tekstov (Machine Translation Based on Semantic Coding of Texts). — Nauchno-tekhnicheskaya Informatsiya, Vol. 1. 1963.
16. Yaglom, I. M. and A. M. Yaglom. Veroyatnost' i informatsiya (Probability and Information). — Fizmatgiz. 1960.

Chapter V

RATES OF DEVELOPMENT OF CIVILIZATIONS AND THEIR FORECASTING

§ 1. THE IMPORTANCE OF THE PROBLEM OF RATES OF DEVELOPMENT

The existing scientific notions on the possible rates of development of life and civilization in other planetary systems may greatly influence the attempts to establish interstellar communication.

These scientific notions directly determine the estimates of the abundance of life and intelligence in the Universe and the number of civilizations which are technologically more advanced than our terrestrial civilization and, in particular, are capable of sending radio messages over thousands, millions, and billions of light years.

The first question to consider is to what extent the historical rates of development of the Earth civilization (and the rates of its biological prehistory) are typical of civilizations in general. For example, were we to accept the suggestion that the Earth conditions stimulated the biological evolution to a greater extent than the conditions prevailing on other potentially life-sustaining planets, we would be led to regard the Earth civilization as one of the most advanced in our part of the Galaxy. Some evidence supporting this idea will be given in the following. In addition to the evolutionary factors, we also have to consider the exact time at which life originated and the extent to which a given set of conditions fixes the evolution rates (in principle, for identical initial conditions and environmental factors, quantum fluctuations may alter the time of appearance of significant mutations and thus change the timetable and the general course of evolution).

Prediction of the future rate of development of the Earth civilization is also important for establishing radio communication with distant civilizations. Different forecasts of our future ability to transmit and receive messages may lead to different approaches to the entire aspect of the allowed expenditure of capital and effort in this field. These approaches may be guided by the ideas of the internal determinism of the rates and directions of development (i.e., to what extent these growth rates will increase following the reception of signals from more advanced civilizations) and by the importance we attach to maintaining consistently high rates of our scientific-technical progress.

Scientists and science-fiction writers generally assume a great variety of life forms and civilizations on different planets, but tacitly imply that the

variance in the rates of evolution is not pronounced. Thus, Macgovan and Ordway /1/ give three different distribution curves for the length of time between the inception of life and the appearance of civilization.

In each of these distribution curves, the rms deviation is 10—20% of the mean duration. Therefore, if we take 3 to 4 billion years for the mean duration, the evolution of intelligent life on any planet will take no less than 1 billion and no more than 6 billion years. This approach a priori deprives the Universe of some of its inherent richness and variety.

Even on the Earth, the rates of evolution vary between considerable limits in different zoogeographical zones: small isolated zones (Australia, Madagascar) are still inhabited by archaic life forms. Earth analogies suggest significantly slower rates of evolution on planets of the size of Mars or the Moon or on planets with small dry-land area. By the same token, the evolution is faster on larger planets, since all other conditions being equal, the population there includes a greater number of individuals.

The rate of evolution should drop on a planet where the climatic conditions are steadily uniform (no glacial periods) or, conversely, if the climatic eras alternate with excessive frequency. If the equatorial plane of the equator is not inclined to the orbital plane, the tasks of the evolution are made much easier: many of the survival problems are eliminated and the evolution may become more sluggish or even stop altogether. At the other extreme, if the equatorial plane is perpendicular to the orbital plane, numerous life forms will have to migrate over large distance annually, so that all boundaries between individual zoogeographical zones will be obliterated and the probability of branching in the course of evolution will be reduced. The number of large satellites of a planet determines the pattern of nocturnal illumination and thus influences the behavioral flexibility of nocturnal animals. This reasoning applies when the organic world of a planet is not inherently different from that of the Earth, especially with regard to ecology, hereditary mechanisms, and variability of species. It nevertheless seems that the rates of evolution of life forms and social structures — civilizations — may vary between wide limits. Unfortunately, we still have no reliable information on the rates of evolution and development of extraterrestrial civilizations /2/. The analysis that follows will be based entirely on the rates of development of our terrestrial civilization.

§ 2. THE ASPECTS OF DEVELOPMENT OF CIVILIZATIONS

Although such concepts as acquisition, processing, storage and transmission of information are useful in describing the development of a society as a whole, they are insufficient for describing any particular stage of this development. The rates of development of a civilization have to be treated in terms of sociopolitical and economic development, evolution of language and art, development of science and technology, the role of religion, etc. We cannot maintain, however, that these aspects of civilization as we understand them now will apply indefinitely to describe

the progress of the Earth civilization. Moreover, there is no a priori justification for extending these concepts to other civilizations, thus constraining them to follow approximately the same evolutionary course.

Can we be certain, for example, that the stage of religious awareness is equally prominent in all civilizations? On the one hand, there are indications that the development of religion is associated with local terrestrial factors /3/, but on the other hand, even elephants are endowed with religious sentiments and prayers /4/; if this is indeed so, religion is a universal phenomenon. A study of the history of religion for purposes of the general theory of rates of development of civilizations is of particular interest because religion (especially at the later stages of its evolution) is a clear example of a retarding force slowing down the growth of civilization.

Another interesting factor to consider are the relative rates of development of art and science. The discussion revolving around the topic of "physicists and poets" which figured prominently in the Soviet press at the end of the 1950s helped to formulate some questions, without answering them. Feinberg /5/ noted that the rates of growth of science have long since overtaken the rates of growth of the arts and humanities, and the current trend in all probability can be extrapolated into the future. A point which is not so clear concerns the relative significance and value of sciences and arts in the life of a society and its individual members. Our understanding of the laws governing the contemporary evolution of art is still far too fragmentary to be applied in quantitative reasoning.

The development of language and other means of communication between the members of society is another important aspect contributing to a complete description of the evolution of a civilization.

Language and communication

Social intercourse between the individual members of a population, originally the individual animals in a herd, brought about the development of special systems of conventional symbols, in particular vocal systems, long before the appearance of man on Earth. The main distinctive feature differentiating human language from the "language" of animals is its inherent flexibility, permitting introduction of new symbols whose meaning can be explained using only the means provided by the language itself /6/. Note that despite the acceleration in the rates of development of various aspects of civilization, the languages have developed over the last millennia at an approximately constant rate, retaining some 85% of the vocabulary for 1000 years /7/. The vocabulary started growing at a somewhat faster rate during the last centuries and decades, mainly because of the enhanced activity and the advances in science. The vigorous growth of civilization, however, emerges most clearly from the new systems of symbols that have been put into routine use: road and river traffic signs, chemical formulae, the notation of algebra and calculus, theory of sets and mathematical logic, library classification codes and codes for the classification of standards and patents. The slow evolution of language is counterweighed by the rapid change in the proportion of spoken and written words, the radical change in the place of formulae and drawings in written communications, the prominence of slides and movie films in verbal communications.

The conventional language does not even try to compete with the language of formulae and equations. As an example, let us compare the meaning of the word hardness in the following two propositions: gypsum has a hardness of 2, and alloys of this kind are distinguished by their hardness.

The possibility of distinguishing between greatly dissimilar logical meanings of the same word is provided by the context, the semantic redundancy of our speech. Similar multivalued meanings are characteristic of many other words, such as qualitative, height, capacity, etc. It is only seldom that the speaker uses a qualifying word, such as degree, magnitude, amount, or measure, or develops a composite word by adding, say, the morpheme -grade (German). It seems that the "words" to be used in communication between distant civilizations will not necessarily be the words of contemporary languages.

What are the quantitative characteristics of communication that reflect the level of civilization and the rate of its development? These quantitative characteristics are not to be sought in the structure of every individual language, but rather in the number of different symbolic systems which remain after maximum unification and standardization (the number of non-redundant symbolic systems, so to say). These include the statistical characteristics of the vocabulary /8/ and primarily the total number of terms in current use. The scientific and technical progress is even more clearly reflected in the statistical parameters of numbers which occur in published texts. In the course of history, the range of orders of magnitude covered by the printed numbers increased (one-digit, two-digit, and other numbers), and the exponential notation (with a decimal base) is now being used with increasing frequency. The rounding off of approximate numbers is now better motivated, and it is only seldom now that numbers are truncated at units. With the spectacular growth in orders of magnitude, the power exponent is now often rounded off, and one may prefer writing 10^{60} rather than 10^{59}. These tendencies are fundamentally simple, they can be studied without substantial material expenditure, and they are associated with highly general and basic principles. There seems to be no reason why these tendencies should not be generalized to extraterrestrial civilizations.

The relative proportion of the exact (discrete) and approximate ("continuous") numbers in published texts follows a more complex evolutionary pattern. Medieval texts reveal a predominance of discrete numbers: the number of people, objects, operations; the figures of celestial bodies and their orbits are regarded as ideal spheres and circles. The numbers 7 and 3 are reported more often than they occur in reality. The age of enlightenment and the development of capitalist society led to the rejection of some superstitions, substituting measurements for counting and philosophizing. The attention of the engineers was focused on physical processes, primarily those involving power and energy. The discrete numbers no longer filled the foreground. But at the end of the 19th century, the physicists suddenly rediscovered the significance of discrete quantities: the elementary charge was discovered, and then the microcosmos was quantized. Soon after that, both the life sciences and technology switched the main emphasis from energy processes to information concepts. Morphological descriptions were often supplemented by structural characteristics, and the "discrete" mathematics received a great impetus. The

extraterrestrial civilizations will apparently follow a similar course of development.

Conversely, the great variety of the living languages on our planet and the absence of any obvious relationship between the existing language families can be attributed to the peculiar features of the Earth topography, difficulties of sea voyage, etc. On a planet with a more compact dry-land area, fewer mountains and deserts, and a greater abundance of navigable rivers, a single universal language may develop at the dawn of civilization Another sufficient condition for the development of a common language would be the availability of excellent means of transportation or communication.

There are no reasons to believe that a civilization is incapable of creating an artificial universal or international language. Why did this not take place on Earth? Latin was the international language of scientists and theologians in the Middle Ages, but it was very difficult to master and far removed from any of the living languages. Had this not been the case, Latin would have probably remained in international usage and eventually gained a special universal standing. In the period of the industrial revolution, science, technology, and the arts developed in a number of leading countries, of which no two had a common language or any similarity in their languages. This may be the reason why not a single language gained acceptance as a universal means of communication. In the second half of the 19th century, some artificial languages were first proposed. The expectations were very high (Esperanto was considered especially promising),* but World War I only enhanced the nationalistic barriers and mankind was too busy to concentrate on the problem of universal communication. At present, the problem has lost much of its pressing importance in view of the promising prospects of machine translation, the great advances in methods of teaching of foreign languages, and the generally widespread education.

On the other hand, we do not see why a jumble of dialects, languages, and language families even greater than on Earth need constitute a barrier against the expansion of a civilization into outer space.

Demographic characteristics of civilization

The statistical data on population growth are fairly accurate only for the last few centuries. Sufficient data on the proportions of urban and rural population, distribution according to age groups, etc., in most countries are available for the last 150 or 200 years only.

The extensive demographic literature either concentrates on the study and the forecasting of the growth and variation of composition of the population in individual countries or restricts the treatment to relatively short periods. Authors analyzing the growth of the Earth population over the last centuries either do not consider the topic of fitting all the available data with a single analytical dependence or arbitrarily assume an exponential function with a correspondingly low accuracy. And yet, the Earth population data closely follow a hyperbolic function of time, as has been shown by Shklovskii /10/. Shklovskii has also noted, however, that this

* There are indications that Esperanto is used as a living language in some rural settlements in British Guiana /9/.

hyperbolic dependence will soon break down as we learn to control society to a greater extent. The increase of the percentage annual growth may slow down also because of widespread automation and increased mechanization, which will make the economic ceiling of a country less rigidly defined than it was before. Rapid economic growth necessitates an advanced socioeconomic structure, proper organization and planning of industry, education, and last, but not least, a sufficiently high rate of population growth.

The interpretation of various forecasts of the future population of the Earth may meet with certain difficulties because of the possible discovery of life forms on the threshold of intelligence /11/ and also because of the possible creation of near-intelligent systems /2/, teaching of dolphins /12/, breeding of normal animal species with greater than normal brain capacity through surgical intervention or selection, or development of man-simulating machine programs. Forecasting difficulties of another kind are associated with the future possibilities of suspended animation. All these unaccountable factors naturally introduce a considerable error in the determination of the Earth population.

In application to other civilizations, the concept of population is even less certain because of the inherent difficulties in the definition of the concept "organism." For example, in application to the bees and ants of our world, some authors treat the entire beehive or anthill as an organism, rather than a genetic individual /13/.

Social and cultural changes in society may disrupt the quantitative rules of growth of the entire population and of separate demographic indices. Thus, the percentage of the human population which inhabited cities with a population of over 10^5 varied as follows in different years t /14/:

t	1800	1850	1900	1950	1960
p	1.7	2.3	5.5	13.1	20.1

It is readily seen that the growth of this percentage p of the population, and especially the growth of the ratio of the urban-to-rural population $p/(100-p)$ does not follow a geometrical progression: the relative growth rates are distinctly accelerated here.

The data of this table can be approximated with an equation of the form

$$p\% = \frac{110}{110 + (2000 - t)^{5/3}} \cdot 100\%.$$

This model can be interpreted as showing a tendency of the entire Earth (or each continent) to develop into a single large city toward the end of the 20th century. On the other hand, the current tendencies point to a definite deurbanization, with the urban population migrating to the suburbs. It would seem that large-scale development of television, a single telephone and videophone network, remote access to libraries, etc., will retard the growth of large cities.

The development of individual abilities

The average (or the record) intellectual ability of the individuals and the extent to which this ability is utilized also provides a certain criterion

for evaluating the level and the rates of growth of a civilization. This analysis, however, involves considerable difficulties, mainly due to lack of objective and reproducible means for measuring individual ability and assessing the coverage of these measurements and the choice of significant (and independent) characteristics from the entire set of alternatives.

An index which is particularly convenient for measurements may prove to be ineffective for estimating the growth dynamics if we are unable to determine its value for the past epochs. So far, no effective methods have been developed for measuring the general intellectual abilities of men. The widespread IQ tests yield a numerical scalar index which lumps together the inherent creative abilities, the sum total of knowledge and experience, and the cultural level of the person tested. Psychologists are now ready to admit that these tests are deficient.

In their efforts to pick out the gifted, as well as the knowledgeable, students, the teachers in some colleges and universities have developed a certain diagnostic power (intuitive, rather than scientific) which enables them to identify outstanding creative ability through various educational competitions.

This intuitive experience, however, is very difficult to apply to estimating the abilities of the scientists or inventors of the past, when the sum total of our knowledge, the general picture of the world, and the teaching methods were entirely different and the people faced problems of a completely different nature. The present-day population of the Earth is many times larger than the population in ancient times or in the Middle Ages, education is much more readily accessible than, say, in Ancient Greece, Is it not natural to assume that among the modern scientists there are minds which are at least one order of magnitude more brilliant than Archimedes, Leibnitz, Newton, and Lomonosov? Is it not likely that Bohr, Wiener, or Landau would have achieved much more in Descartes' place? Similar questions can be raised regarding other fields of human endeavor: poetry (Pushkin or Blok), prediction of future technology (Roger Bacon or Wells), chess (Morphy or Alekhin). We can hardly endeavor to answer these questions without far-reaching advances in the psychology of creative abilities and in the psychology of constructive education, without going in minute detail into the particular problems that science and culture faced in every epoch, and without analyzing the possibilities that each period presented to people. It is very difficult to find problems which are of the same difficulty in different periods. The number of foreign languages that can be learned depends not only on the mind but also on the memory; another factor to remember is that in different epochs, foreign languages occupied positions of different importance in elementary education.

There are very few creative problems for which independent solutions kept cropping up over the ages. The lost proof of Fermat's Great Theorem could not be reconstructed over three centuries, but how are we to be sure that the original proof did not contain an error?

Perhaps civilization can be measured in terms of individual achievements, the extent to which the mental and the physical potential of the human organism is utilized. Here again, much is still unmeasurable. For some professions, the productivity of labor has been thoroughly documented over the ages, but how are we to assess the accomplishments of a military commander, a teacher, a science writer? Chess masters can be graded according to the depth and the complexity of the combinations

that they discovered (or missed), but we have no means for estimating to what extent their ingenuity has been aided by the sum total of the historically accumulated experience and knowhow.

The various sports are in a more advantageous position in this respect. But nevertheless, comparison of distant epochs involves difficulties. The various records are kept only starting with the 19th century; they are available with high accuracy with full description of the rigidly controlled conditions. These conditions, however, did not remain constant either: the footwear and the starting conditions have changed for the sprinter, the pole-vaulting pole is now made of a different material, etc. The number of new sports increases rapidly, and there are correspondingly fewer sportsmen specializing in every given branch (in proportion to the total). On the other hand, the specialization and the strict professional approach somewhat increase.

We have briefly considered some aspects of the growth of civilization which clearly show the difficulties associated with estimating the growth rates, the variation of growth rates, and establishing quantitative expressions for the relevant regularities. We are in a much better position, however, with regard to the rates of development of the other aspects of civilization, such as economy, technology, and science. These aspects are of the utmost importance for elucidating the possibilities of space travel.

In the next two sections, we will consider in greater detail the technological and the scientific aspects of civilization, but again we will only present a qualitative description: at this stage, we are more concerned with the overall picture of the accelerated rate of growth, the conditions under which the relevant indices describing the development change, and other topics of this kind. Quantitative characteristics are still available for relatively small time periods, and it is not at all safe to generalize them to extraterrestrial civilizations.

§ 3. INDICES OF TECHNICAL PROGRESS

One of the principal characteristics of a civilization is the level of technical knowledge, the indices of various technological means, the quantity and the quality of manufactured products, the amount of energy used, etc. Another important factor is the proper organization and comprehensiveness of the agencies of control governing the entire complex of technical means and the utilization of technology as a whole.

Science, medicine, and education are also largely dependent on the level of technological knowledge. At present, however, technology is no longer merely a means for scientific research, but actually a prime mover in the advancement of science, constantly challenging scientists with new problems and tasks.

Extensive literature is currently available on the economic, technical, and scientific history of individual countries and humanity as a whole. Numerous studies have been devoted to the growth of individual development indices (power, velocity, etc. /15, 16/). This information, however, mainly refers to the last decades. The information for the last centuries is substantially less comprehensive.

Comparison of numerical data for various epochs of technological development is not equally valid for all the indices: the growth of energy output is a more appropriate index than the growth of the pool of metal-machining lathes, since the productivity and the precision of a lathe markedly changes with technological development; the capacity of transport expressed in ton-kilometers is a better index than the number of seagoing vessels. In general, more valid conclusions can be based on growth indices whose unit retains a constant value over the various epochs.

The production growth indices, which have been traced over a long period of time and remain comparable over the entire period, include the world-wide production of silver or coal, whose rates steadily increased starting in the 16th or 17th century (when the first more or less accurate quantitative data were recorded) and up to the end of the 19th century, and then somewhat slowed down. However, coal is now largely replaced by petroleum, and the rates of growth of petroleum production are much higher than the present (and past) rates of growth of coal production. Similarly, silver has been partly replaced by other chemically resistant materials (including plastics), whose production now increases at rates which were unimaginable for silver.

In the development of transport, the speed and the power of one form inevitably reach a certain limit, and then that form of transport is replaced by a new form which continues developing, and so on.

We thus see that the rates of growth of a civilization are characterized by a succession of changing leading indices.

On the succession of indices

The above examples show that each stage of development of a civilization is characterized by certain basic indices which are replaced by other indices at later stages. Moreover, we can speak of a certain succession of indices between the evolution of the animal world and the development of human society. The development of human society is governed by a characteristic acceleration of growth rates (and changing of indices) which began in the course of the biological evolution /17/. It is even possible to set up an evolutionary scale from $5 \cdot 10^9$ years to 50 years, with the successive periods diminishing by a factor of 10, and compare the principal stages in the development of life, society, and technology to the divisions of this scale /18/.

When we pick out a significant index from a number of succeeding technological means, we are never sure that this index will retain its significance in the future also.

In the Middle Ages, considerable emphasis was laid on the development of thermally insulating materials and on the strength of materials in bulk, whereas later the focus shifted to electrical insulators and the strength per unit weight (as man progressed from the building of fortresses to the building of skyscrapers).

Not only the relative role of the growth of various indices in a given direction of technological progress changes with time, but the relative importance of the different directions of progress is also variable. As

an example, we can mention the advent of computers and control systems, which are gradually replacing the power systems as "machines" in the foreground of technology. On the other hand, the power and energy line stretches through the entire history of life, and not only technology, on our planet.

We could continue this extrapolation of ever increasing scales of activity which engender the succession of the leading (in terms of speed or significance) aspects of civilization and possibly govern the succession of periods when the rates of growth are of primary importance for survival (or in general of relatively high importance) and periods when they are insignificant (or of relatively low significance). At present, however, science does not have sufficiently reliable tools for measuring the level of the principal aspects of civilization, the rates of growth, and the significance of their development. Not even the exact function of each aspect at every stage of development of civilization has been established. We are not at all certain, for instance, that prehistoric religion fulfilled any positive function, not even mnemonic, helping to assimilate and retain in our memory the great variety of empirical factors by dressing them up in a digestible coat of legend and superstition.

If the growth dynamics of every individual technological means is nearly exponential, i.e., the percentage annual growth of the corresponding index is constant, then for an index covering a number of succeeding technological means, we should take into consideration the change in the absolute amount of the annual growth. Without special "scaling" of the time factor in accordance with the general rates of development of technology and accumulation of technical information, we will simply be unable to grasp the multitude of numerical data relating to the development of different branches of technology in different epochs.

Although medicine and education have some features in common with technology, their growth rate and progress is much more difficult to gauge and measure. Some of the reasons follow.

The achievements of medicine can be assessed in terms of the mortality index, the general state of health of the population, and its ablebodiedness. The state of health and ablebodiedness, however, are not easily expressed by an objective and reproducible numerical index; moreover, these factors depend not only on medicine, but also on the socioeconomic conditions, work and leisure conditions, hereditary predilections of people who have reached a certain age, etc. It is very difficult to allow for the fact that the life expectancy of people with some pathological hereditary defects nowadays is not as different from the life expectancy of normal people as it was some time in the past.

The difficulties in measuring the progress in the effectiveness and quality of education are associated with the great diversity of the effects of education. Education means knowledge, both applied and abstract, various "skills," including the skill to apply the acquired knowledge and to acquire new knowledge. It also means the cultivation of inbred abilities and the results of training. Social, economic, and technical changes in the life of society have a prominent effect on the development of children in the school stage, as well as on the mind of a specialist after graduation.

Mathematical functions describing growth rates

We do not give any actual numerical data mainly because the exact figures relating to the growth of silver production in the world or the progress in oceanography are of no particular relevance from the point of view of the search for extraterrestrial civilizations. The rates of growth of energy and power output have been discussed in Chapters I and III. No general conclusions can be drawn regarding the rate of development of radio engineering, since the period in question is obviously too short. This reservation is even more applicable to the technology of space flight. The main purpose of this section is to throw light on the change and succession of the leading technological indices. Therefore, in practice, we cannot say in what terms we should characterize the technology of the supercivilization with which we hope to establish a communication, and what indices we are to apply to describe its level and rate of growth.

Nevertheless, although we do not intend to give numerical characteristics of the rates of growth of civilizations, we can offer some comments regarding the mathematical expression of these rates.

A constant growth rate corresponds to a linear dependence of the corresponding index on time. This is very seldom the case for the leading indices. Most indices characterizing the development of civilizations display rapidly accelerating growth rates. If the growth rate is proportional to the value of the index (i.e., the relative growth rate is constant), the corresponding index increases exponentially with time.

Finally, if the relative growth velocity also increases, i.e., the rate of change increases sufficiently rapidly with any incremental change in the index, the index is seen to grow hyperbolically. A characteristic feature of the hyperbolic law of growth is that the index will rise to infinity in a finite period of time. Note that numerous indices of growth and development of the Earth civilization are adequately fitted with the hyperbolic curve (e.g., the demographic index). However, our remarks regarding the change and succession of the indices indicate that the hyperbolic phase eventually breaks down for every index.

In a number of cases, the parameters characterizing the development of civilizations have some intrinsic restrictions (e.g., the ratio of the number of scientists to the total population). The ratio of growth of this index naturally increases as it approaches its natural limit. Let n denote such an index, and suppose that in the early stage of development it follows the regular exponential dependence

$$\frac{dn}{dt} = \text{const}\, n. \tag{5.1}$$

At later stages, as the value of the index approaches the limit, the rate of growth will slow down. We may thus assume that $\frac{1}{n}\frac{dn}{dt}$ will be proportional to the difference between the maximum value of the index (e.g., 100%) and the value at any given time, i.e.,

$$\frac{dn}{dt} = \text{const}\,(n_{\max} - n)n. \tag{5.2}$$

Or, in a different form,

$$\frac{dn}{dt} = \frac{n}{t_0}\left(1 - \frac{n}{n_{\max}}\right), \tag{5.3}$$

where t_0 is the time scale. The solution of this equation gives the so-called "logistic" curve

$$\frac{n}{n_0} = \frac{n_{\max}}{n_0 + (n_{\max} - n_0)\, e^{-t/t_0}}, \tag{5.4}$$

where n_0 is the initial value of the particular index (for $t = 0$). The characteristic feature of the logistic curve is that it is symmetric about the inflection point. If the curve is not symmetric, it is not a logistic curve /20/.

Other curves are also used for similar purposes. For example, the hyperbola can be replaced by Zeman's equation /19/

$$n = \text{const} \lg \frac{\text{const}}{t_0 - t}, \tag{5.5}$$

which also leads to accelerated relative rates, but ensures a finite value of the corresponding integral (i.e., the total number of "events" remains finite).

In exponential variation, the index always takes a fixed length of time to double its value. It is for this reason that the growth rates are often characterized by the time to double the value of the index /20, 21/ (see also next section). For other curves, however, the time of doubling is not an invariant.

§ 4. RATES OF GROWTH OF SCIENCE

Science is one of the principal aspects of civilization whose importance steadily and very rapidly increases. The normal activity of a lathe operator, or a doctor, gives results which are limited to the immediate neighborhood. The activity of a scientist, on the other hand, may benefit the whole of humanity.

On the other hand, ten identical parts turned out by ten lathe operators are ten times as valuable as a single part. Conversely, ten identical research projects undertaken by ten scientists are hardly any more valuable than one of these researches. Science is intrinsically different from material production, medicine, education, etc., in that, first, each and every one of its products should be brand new and, second, any product of science is not subject to wear and tear under any circumstances, it does not require maintenance and repair and it can be reproduced at any time in any quantities. A scientific fact can be used simultaneously in several places around the world, whereas a lathe is fixed in its location. The product of science is not a simple additive sum of all the resources: its effects are definitely multiplicative in that they themselves act to increase the resources of all mankind. Another difference between science and technology is that a scientific product becomes common property, properly

mechanized and automated, at a substantially later stage of development of society than an industrial product. The basic requirement of novelty imposed on every scientific product enhances the importance of chance and accident in the process of acquisition of scientific knowledge as compared to the more orderly and systematic nature of material industry and medicine.

Because of these intrinsic differences, science has to be treated separately from technology, although science is currently gaining in importance as an independent productive force and although scientific research is often oriented to the immediate needs of industry. The qualitative characteristics of the growth rate of science therefore merit a section to themselves.

The main problem in measuring the growth rates of science is the choice of the significant indices. The number of scientists engaged in research and the financial allocations do not provide an accurate picture of the role of science in society, although these parameters increase very rapidly with the development of civilization.

Scientific knowledge is the main product of scientific research, and the growth of this knowledge is a basic criterion of the advancement of science. However, it is very difficult to form an objective estimate of the amount of knowledge, and as the main working parameter one generally uses the number of research workers, the number of scientific publications and reports.

During the last three hundred years, these indices increase on the average following an exponential curve /20/, but the time of doubling is different in different branches of science. In physics, the total number of publications is doubled in about 10—15 years /20/, whereas in some subdivisions of mathematical statistics the corresponding period has lately been as short as two years /22/. The financial allocations and the relative number of research workers also increase exponentially, although eventually this parameter will describe a logistic curve.

The number of creative workers and the proportion of time devoted to creative and noncreative work by scientists is also variable with time /28/. Price estimates that the number of really creative scientists is roughly equal to the square root of the total number of research workers, and it is they who author approximately half of all the publications and 70—80% of the significant results.

At this stage, we can assess the amount of valuable and significant discoveries in different epochs only by intuition. These intuitive estimates, although highly subjective, try to guess the number of significant stages in science. Feinberg /5/ is of the opinion that the scientific discoveries of our century are of essentially the same relative significance as the scientific discoveries of each of the last three centuries. It is in this way that he defends the hypothesis of the exponential growth of science. He points out that only the absolute increment of scientific knowledge, or scientific cognizance of the world, increases, and that the growth of this absolute increment, combined with the fact that in our century science has finally caught up with and overtaken the role of the arts and humanities in our society, are responsible for the false impression of the ever increasing significance of each successive decade, of each successive century, in the shaping of our scientific knowledge and the scientific picture of the Universe.

However, Feinberg's examples are borrowed mainly from the field of physics, whose development began earlier than the other sciences; many of the problems of the microcosmos were solved in the previous centuries. Physics probably is not the best example for judging the projected rates of progress of science in general. The significance relations of various five-decade periods in the history of astronomy and geophysics, chemistry and biochemistry, physiology and genetics look quite different.

Similar arguments, however, apply to any subjective line of reasoning. Let us consider the actual growth of the number of scientific discoveries in some fields of science. The very choice of significant discoveries is highly subjective, but, once the choice is made, the data can be processed by fairly reproducible methods. Let us trace the frequency of occurrence of the dates of various epochs in review monographs dealing with the history of some branches of science.

The table below gives the frequency of references to various centuries in one of the fairly popular books on the history of mathematics /25/. Dates of birth and death of the individual scientists and dates relating to the development of the history of mathematics (as distinct from the development of the science of mathematics) were not counted; dates mentioned in footnotes and in verbal form (not numerical) were also ignored.

Period (centuries)	B. C.	1st–14th	15th	16th	17th	18th	19th	20th
Number of dates mentioned	30	50	9	19	79	119	255	9
Possible model			8	20	50	125	312.5	

The growth from the 15th to the 19th century was close to a geometrical progression, apart from the exceptional 17th century, the highlight period in the history of mathematics. The significance of the mathematical discoveries of the last decades naturally has not received a full evaluation and we can hardly expect them to appear in a popular book. It is significant, however, that there are 139 references to the first half of the 19th century, and only 116 references to the second half. Exponential growth of the number of dates persists only up to the first half of the 19th century.

Let us now consider the frequency of dates in a source book in the history of psychology /26/. The dates start back in the fourth century B. C. and the frequency of references increases steadily up to the early 1930s. Identifying the years t_γ in which the cumulative number of references reached an integral power of 2, i.e., 2^γ, we obtain the following table (rounded off to whole decades):

γ	4	5	6	7	8	9
t_γ	1580	1670	1780	1850	1890	1930
Doubling time	90	110	70	40	40	

The chronological table of the important discoveries and inventions in the field of chemistry /27/ give the following years as the dates in which the cumulative number of events reached an integral power of 2 (double

dates, originallly written with a dash, are replaced by the arithmetic mean):

ν	2	3	4	5	6	7	8	9	10
	50th century B.C.	22nd century B.C.	1st century B.C.	15th century	1630	1780	1820	1860	1915
Doubling time	3000	2000	1500	200	150	40	40	55	

The last two tables shows that the cumulative number of references has been growing with a more or less constant doubling time for the last 150 or 200 years only. On the whole, even the progress of the last three-four centuries cannot be fitted with a single exponential curve. In some books the cumulative number of references is considerably slowed down starting as early as the 18th century. This trend is most pronounced in one of the books on biology /28/. In a number of monographs on the history of science, the number of references according to centuries or smaller units of time increases at a nonuniform rate: for example, the history of linguistics /29/ shows a sudden upsurge in the number of references in the 16th century, with smaller bursts in the 13th century and the last quarter of the 19th century.

Thus, our attempts to estimate the volume of knowledge and the productivity of scientific research at different stages of the development of science or individual branches of science fail to detect sufficiently objective and yet significant indices. The number of pages published in journals is too superficial an indicator, and the number of discoveries is too subjective.

In each field of science there are probably more reliable and objective data, such as the measure of completeness, accuracy, and reliability of the scientific knowledge. A measure of completeness is provided by the ratio of the number of studied objects of a certain class to the number of objects of the same class which have not been studied; a measure of accuracy is provided by the number of significant digits in the results of measurements of certain parameters, and as a measure of reliability we may use the length of time needed to detect insufficiencies in the accuracy of certain parameters. The search for objective indices of the development of science is still at its embryonic stage.

The subdivision of the aspects of human civilization presented in the last sections follows the traditional line of reasoning, and is by no means the best for our purposes; we would be better off operating with more general terms of stationarity, determinism or regularity in flow and storage of information, energy, matter, etc. However, a half-baked transition to a new system of concepts is no more advisable than any attempt to formulate the final conclusions in traditional terms, which are intrinsically suitable only for the solution of traditional problems covering much less ground than the problems discussed in this book. At the present stage of the work on the problem of progress rates and forecasting in relation to the entire topic of communication with extraterrestrial civilizations, we are faced with a necessity of collecting a large volume of various facts, verifying them, and reclassifying on a new basis. Any attempt to arrive at a

definitive assessment of the interpretation of the facts or even of fact selection from the standpoint of a particular, restricted conception or a particular branch of science, whether thermodynamics or semiotics, will only obstruct future effective approaches to this entirely new and unusual problem.

§5. FORECASTING

Control systems in nature, industry, and society are equipped for information acquisition and are capable of classifying the outside stimuli from the point of view of the required system response, which is intended to ensure preservation and possibly development of the system itself or of some larger cybernetic system /30—32/. With some reservations, we can also discuss output of information from the system. Simple systems evaluate these outside stimuli only in order to determine the state of the internal and the external media at the material time, whereas more complex systems can respond to a forecast future state of the environment as predicted on the basis of the current measurements. This extrapolatory or forecasting function of control systems has recently attracted considerable attention in biology /33/ and in engineering cybernetics /34/.

The importance of forecasts increases as human society reaches progressively higher levels of complexity and civilization develops. Forecasting, in the form of prophecies, was one of the functions of the ancient tribal chief. Professional oracles and prophets were a common phenomenon in ancient society, and the truthfulness of their prophecies (e.g., prediction of eclipses) is often attributed to empirical knowledge. The social recognition enjoyed for a long time by various oracles, astrologists, palm readers, etc., is associated with their exceptional understanding of human psychology. Intuitive methods of forecasting, whether truly correct or simply interpreted as such by the anticipating customer, are generally erroneously motivated by the configuration of lines on a palm, the position of planets, or the combination of playing cards, etc. However, the first advances in modern science completely undermined the prophet's authority in the educated strata of society. This trend dates back to the theoretical work of the French materialists in the 18th century. Unfortunately, in rejecting the parapsychological techniques, they did not investigate the likelihood of fulfillment of various prophecies relating to the fate of individuals (it is very difficult to analyze this factor, probably because of the suggestive influence that such a prophecy may have on the future fate of the individual), nor did they study the psychological mechanisms of prophecy and forecasting. During the last 200—300 years, the most significant, astonishing and reliable predictions were made by the leading authorities in each field of human activity. A correct forecast of the outcome of a military campaign was expected to come from the military staff or the politicians, a chess master was regarded as the best authority to offer an opinion on the possible outcome of a game. During the last decades, the situation slightly changed because of the closer intercoupling between the various forms of activity and various fields of human knowledge. The emergence from the stage of narrow topical specialization of forecasts and predictions inevitably led to some sort

of a professional specialization and establishment of groups and organizations whose business is to forecast the future in all fields. On the other hand, science-fiction writers occupy a progressively more important role in human thinking. Science-fiction writers have become quite specialized: Jules Verne wrote a considerable number of pure adventure and travel novels, without any vestiges of science fiction, whereas today a respectable science-fiction writer will hardly write a non-science-fiction novel. Outstanding scientists acquire a taste for science fiction: H. G. Wells graduated as a biologist, Isaac Asimov is a well-known biochemist, Arthur Clarke is an astronomer, I. Efremov is a paleontologist. Some science-fiction writers eventually switch from literary treatment of their ideas to systematic analysis of their conception of the future in treatise or monograph form /35, 36/. Subsystems specializing in the forecasting of the future thus again acquire a special position in the fabric of our civilization.

Classification of forecasts

Forecasts can be divided into those dealing with mass events (which recur without significant changes in the relevant conditions) and occasional or unique events (which are very seldom observed, if at all). The mass events are naturally easier to forecast.

There are other possible approaches to the problem of forecasting.

Classification according to the scope of the problem: a) individual or particular forecasts ("I may lose this peon"), b) forecasts relating to significant aspects of the fate of an individual, a group of individuals, a scientific experiment, etc., c) forecasts relating to the development of a certain branch of industry or science, d) forecasts of the development of the entire civilization, e) forecasts purporting to predict the reaction of other civilizations to the reception of an intelligent signal or the discovery of some apparatus launched by the originating civilization.

Forecasts are often classified according to the length of their range or term /21/. In this classification, unfortunately, the unit of time is a year or a century, rather than an epoch defined as the time of doubling of a significant index. The calendar units used as the exclusive basis for these forecasts naturally invalidate all comparison of forecasts prepared by a civilization in periods characterized by different rates of progress or by different civilizations.

The forecasts can be divided into logically sound and intuitive; another division is into forecasts using only qualitative data and those based on both qualitative and quantitative information. This subdivision can be extended to cover the criteria applied in the selection of the experts for the preparation of forecasts: the forecasters can be selected on the basis of some logical tests and criteria or by simple intuition.

In terms of the outcome, forecasts may be deterministic or stochastic; stochastic forecasts may present a discrete probability distribution, a continuous distribution of some numerical variable, a distribution in the function space, etc.

Without going in detail into the systematics of classification of forecasts and the relationships between the various classifications, we will try to consider the means and ways for preparing reliable forecasts and improving their accuracy.

Accuracy of forecasts

The first step is to learn to compare the accuracy or reliability of various forecasts or series of forecasts.

The information deficit that the particular forecast failed to predict or foresee may be used as a quality or reliability criterion for stochastic forecasts. Let the possible discrete outcomes K of the forecast event be assigned the probabilities $P_K > 0$. Suppose that the actual outcome was K_1. Then $-\log P_{K_1}$ defines the information deficit: it is zero for $P_{K_1}=1$ and slowly increases to infinity as P_{K_1} approaches zero.

Comparing different forecasting methods (or systems) $1, 2, \ldots, j$, applied to events $1, 2, \ldots, i, \ldots, n$, we draw up the sum of information deficits for each method: $S_j = -\sum_{i=1}^{n} \log P_{j/K_i}$. S_j provides a reliability criterion of the series of forecasts obtained by method j: the smaller the probabilities assigned to the true outcomes, the poorer is the forecast and the higher is S_j. This criterion is additive and convenient in applications. It is readily generalized to the case of a continuous probability distribution, when the result of checking the forecast is expressed as some approximate value of the unknown.

The above criterion can be used for evaluating various modifications of a forecasting technique, for assessing the qualifications of various experts, different schools of thought, etc. The criterion can also be applied to verbally expressed degrees of certainty in different outcomes of a discrete distribution: the frequency of errors is used to evaluate the probabilities which are hidden behind such expressions as "possible," "unlikely," "impossible," "absolutely impossible."

An expert, having familiarized himself with this error statistics, will eventually be able to derive stable numerical estimates of probabilities from such loose expressions, whereas an inexperienced person may easily interpret these expressions as corresponding to any number between 10% and 0.01% likelihood of outcome.

If society is interested in evaluating the state of its science and establishing to what extent science is capable of assessing the reliability of hypotheses on the basis of indirect evidence, more emphasis should be placed on polling among the members of the scientific community and on detailed tests of "intelligent" machines of various types, starting with basic pattern recognition programs.

It is particularly important to canvass for opinion before the final analysis of the results of critical experiments, observations of fundamentally new phenomena, exploratory trips to new parts of the world, of the planetary system, and other planetary systems. A society keeping a complete record of the various stages of exploration of the outside world will attain a better grasp of its own potential.

Since this has never been a common practice on Earth, and the psychological mechanisms of forecasting have been little studied, we are not in a position to arrive at a precise and comprehensive comparison of different forecasting techniques. The helplessness and the ossified approach of various scientists and science-fiction writers emerges with great clarity when dealing with far-reaching forecasts of the scientific and technological

trends of our civilization /36/. Practical failures often intermix with scientific inadequacies: both in psychology and in cybernetics, very little attention has been paid to the basic problem of formulation of a new hypothesis, as opposed to the selection of the most likely hypotheses from a given set /37/. And yet, the truly creative activity of mankind cannot be effectively fitted within the limited framework of the concept of selection. This in particular probably provides a partial explanation to the consistent failure of computers in problem solution, the failure of "heuristic programming," etc. /38/. To predict in advance the logical likelihood of the invention of computers or lasers, say, one had to have a very wide grasp of sciences and to operate with such abstract concepts as "materials" and "energy," "control," and "information."

Forecasting the rates of scientific and technological progress

This particular form of forecasting is better developed than the forecasting of the trends of progress. Examples of far-reaching forecasting of rates and epochs of scientific and technological progress abound throughout the history of our civilization.

Some of these forecasts fall wide off the mark, but there are nevertheless individual valid predictions. The main reason for failure, on the one hand, is an underestimation of the inherent difficulties of research or invention and, on the other, an equally dangerous underestimation of the accelerated growth of science and technology. For example, some ten or fifteen years ago, scientists did not fully realize the tremendous difficulties of machine translation or information-theoretical interpretation of the history of languages, they grossly underestimated the harmful aftereffects of transplantation of organs or the chemical aftereffects of pesticides. In 1900, H. G. Wells predicted that atomic energy would be harnessed at the beginning of the century /39/. On the other hand, he did not foresee the development of the airplane before the middle of the century. Some seventy or hundred years ago, some technical achievements, which have by now become a daily reality, were considered to be impossible (or possible only after millions of years) /36, 40/.

The simplest method of forecasting the growth of some index is to extrapolate into the future a theoretical function which closely approximates the past empirical growth data. The closeness of the approximation can be assessed, say, by the least squares method. This approach gives widely varying results for different horizontal and vertical scales, and the best scale is apparently that which gives a function with statistically homogeneous fluctuations over the entire relevant period of time.

The situation is very uncertain regarding the choice between different alternative functions when using the least squares method and the determination of the number of variable parameters for each function.

Some authors stick to the exponential function and the logistic curve /20, 21/, while others use explicitly /10, 16/, or implicitly, to judge from the scales of their graphs /17, 41/, hyperbolas, exponentials of exponentials, rational functions, etc. in certain periods of time.

Formal extrapolation of the dynamic series inevitably leads to errors or to considerable uncertainties in the forecast, if we ignore such factors as socioeconomic changes, probable discoveries and inventions, etc. This severely limits the period of validity of the particular forecasts (development of individual technical means, narrow branches of science, etc., cf. §2), whereas the general laws governing the dynamics of rate of growth are preserved over longer periods /20/.

Psychologically, numerous forecasting errors can be attributed to the fact that it is not the epoch which is determined by the rate of growth, but rather the rate of growth is determined by the epoch. A scientist or a science-fiction writer estimated the progress in this century using the data of the last century, and the progress of the current millennium from the data of the previous millennium, etc. However, the last 300—400 years are characterized by a much higher rate of progress than the entire past millennium. There are a number of significant pointers indicating a steady acceleration of the growth rate over the entire duration of modern history. The results of this forecasting approach thus lead to an apparent deceleration of the rate of growth in the future, in that the future rate of growth is a mirror reflection of the past growth rates. Consequently, the forecast dates and epochs will considerably lag behind the actual accelerated development of science and technology. Thus, the symmetrical mirror extension of the sequence of dates 1500, 1800, 1900, 1950 into the future is the sequence 1950, 2000, 2100, 2400, whereas the present acceleration of the growth rates gives an extrapolated sequence 1950, 1980, 2000, 2015.

There is no evidence to suggest that this last extrapolation is correct, but the former extrapolation constitutes an extreme case of subjectivism whereby the current epoch is adopted by the author as the center of symmetry of the time growth curve.

This subjectivism emerges already on the cover of A. Clarke's book /36/, where the dates 1800, 1900, 1950, 2000, and 2100 are written one under the other. The critique of conservative forecasts in this book is restricted to the psychological level and the level of the actual past history of technology. Clarke's own forecast of the future growth of technology, although not accelerated, does not reveal a special accelerated trend. However, the conception of contracting doubling times is definitely reflected in the forecast growth rates. And yet the currently available future forecasts (see, e.g., /40/) deal mainly with the prediction of dates and epochs, as does Clarke's forecast, and not with the safe growth rates.

The leading importance of the rates of development in any logical forecast was correctly emphasized by Stine /16/, who unfortunately was carried away by formal extrapolation, without meaningful analysis. During the six years after the publication of his work, many of the material growth indices slowed down (e.g., the increase of transportation speeds). Their hyperbolic growth gave way to exponential or even slower rates. This correction, however, does not affect Stine's basic idea, namely that the rates of growth predicted by science fiction fall short of the real growth rates: the growth curve of science-fiction writers is probably concave from below, whereas Stine assumes a straight line.

The forecasting of growth rates of a civilization which has emerged into outer space, populates a certain part of the galaxy, and continues developing rapidly encounters specific difficulties associated with the fact

that no information can propagate at velocities faster than the velocity of light, so that there will be a considerable delay in exchange of information and communication between distant parts of the civilization. In this case, the communication between the different parts of this galactic civilization is not unlike the communication between entirely different civilizations.

Forecasting the growth rates of the Earth civilization

Let us now consider some of the topical factors which have bearing on the forecasts of the future development of humanity. First note that the forecast growth rates (expressed by some mean curve, rather than a whole family of curves with probability measures marked for each) are generally determined for normal conditions: there will be no nuclear war, no lethal microbes will be imported from other planets, etc. The probability of such a global holocaust is estimated by some authors to be currently higher than in the past centuries (when there were instances, if not of total destruction, then at least of a substantial slow-down in the growth of a civilization, e.g., the fall of the Roman Empire). However, even if we accept that the probability of a catastrophe has indeed become higher for a particular year or for the life span of an individual, it does not mean that this probability is significantly higher for the entire epoch, since the length of the successive epochs has shrunk considerably.

On the other hand, the probability of all mankind being wiped out by some natural catastrophe and the probability of destruction due to external forces, as opposed to forces operating from within the civilization, has decreased markedly. We do not foresee a significant danger to mankind as a result of a sudden fall in the level of solar radiation or the explosion of a nearby supernova. Since the probability of such events is vanishingly small, destruction of the civilization due to natural forces is virtually improbable at the present level of our technology, and external and internal factors should not be lumped together, as has been done by some authors /42/.

The most common type of forecasts published in the literature is based on the conception of exponential growth of some index, e.g., power consumption /10/, or of science and technology as a whole /30/. Is this conception borne out by the state of things as we face it now, at the end of the 1960's? There are several significant indications to the contrary.

First, once we assume a certain quantitative dependence of the growth rate, the index of progress need not remain the same all the time: there may be a succession of leading characteristic indices describing the development of the whole civilization or of its individual branches.

Second, as we have seen above, the rate of growth of numerous indices has been accelerating until recently. There is no reason to suppose that this accelerated growth will cease at this particular time. Conversely, it is more logical to assume that we are heading for a number of jumps in the coming years and decades, which will accelerate the rate of development of our civilization even further. After World War II, the leading countries of the world invested enormous means in the design of computers, programmed teaching (and other new teaching techniques), machine translation, elementary particle research, space exploration. All these investments have so far yielded only a minor fraction of the expected returns. As a result, there

are rumors that we are no longer nearing a spectacular jump in these fields. Are these suspicions well founded? After all, space exploration is now progressing at a very fast rate and has already yielded valuable scientific results.

New teaching methods have also proved to be highly effective, although on a limited experimental level only. The problem of machine translation, however, is still far from its solution, but previous research in this field has helped to clear the air and to define the various issues connected with recognition of written messages. Further advances in this field will greatly promote our understanding of the psychic activity of man and thus lead us toward a successful solution of numerous problems of teaching and learning, organization of creative labor and design of "thinking machines."

An important factor in the acceleration of the growth rates of civilization is obviously associated with the social development of our society.

Decoding of messages from extraterrestrial intelligences will naturally also enhance the rate of our progress. The rapid growth of radio astronomy and the projected installation of optical telescopes on the Moon or other atmosphere-less celestial objects greatly increases the probability of reception or interception of such messages in the near future. There is also some hope of detecting traces of technical civilization on the surface of asteroids and satellites, not subjected to weather erosion.

A third objection against the exponential conception is provided by a number of circumstances which are liable to slow down the growth of our civilization in the more distant future (21st or the end of the 20th century). Some of these factors are purely terrestrial /21/: they are associated with difficulties of orientation in the growing torrent of scientific information, the negative effects of the narrow specialization of scientists, etc. Another fundamental reason is the great difference (by a factor of 10^4) in the distances to the outermost planets of the solar system and to the nearest stars. Because of this disparity, some authors think that there will be nothing new to conquer and explore in space for some time after the conquest of the solar system.

A similar situation occurred in the past toward the end of the 19th century, when the white spots had disappeared from the map and yet no technical means were available for deep ocean research. This caused a marked slow-down of the scientific and technical progress, which had accelerated at a very fast rate before, and contributed to the exceptional popularity of the exponential model. When referring to the rapid acceleration of the growth rates before the 20th century, I naturally do not mean the annual acceleration but the acceleration of the periods equal to the doubling time of leading indices (see §4).

It is difficult to foresee how the rates of our progress will be affected by the direct contact with representatives of other civilizations or their automatic machines that is liable to take place in the more distant future. Science-fiction writers advanced a variety of hypotheses regarding the impact of this encounter on the scientific and social advancement of the more backward of the two civilizations. The thesis in /43/ is that no significant change in the rate of progress can be brought about by this intervention from outer space unless the recipient society is to lose its individuality. There is no guarantee that the ideas of enmity and friendship, learning and exchange of information, observation and experiment are universal, and not merely anthropomorphic, and that they reflect the entire gamut of complex

and varied relations between two civilizations. Moreover, once our civilization has found its proper place in an infinite system of interrelated and really friendly civilizations, how are we to be sure that the entire concept of rates of progress will not prove to be anthropomorphic?

If this is really to happen, then when? No one knows whether this will take months or millions of years. Humanity started with the idea of the Earth's unique position in the Universe, and gradually advanced to the conception of multitutdes of inhabited worlds /44/. At the present stage, however, we are all too acutely aware of the abyssal uncertainty on this subject. Some authorities believe in the existence of civilized systems in the Galaxy. Others emphasize that a civilization will hardly need many millions of years to conquer the entire Galaxy, and anyway the time to galactic expansion will definitely be much shorter than the entire history of the Galaxy, so that if an extraterrestrial civilization existed, it is most likely to have appeared long before the origin of the Earth civilization (after all, the probability of two twin civilizations is negligible) and would have by now given signs of its existence. The weak point in this argument is the implicit assumption of the following factors: the rate of development of any civilization cannot be consistently (over many millennia) less than the rate of development of our civilization;* every civilization will be capable of expanding into outer space; every civilization will expand into outer space. These implicit assumptions are a reflection of our anthropomorphic chain of reasoning, and there is definitely no reason to reject the possible existence of other intelligent beings in our Galaxy or in nearby galaxies.

* The assumption of universal growth rates for all civilizations is vividly expressed in /43/: "We are historians, not physicists. We measure time in centuries, not seconds. . . ." In fact, however, history does not retain a fixed unit of time even for our civilization.

Bibliography

1. Macgovan, R. and F. Ordway. Intelligence in the Universe. — N. Y. Prentice Hall. 1966.
2. Schmeck, H. Semi-Artificial Man. — London, G. G. Harrap and Co. 1965.
3. Porshnev, B. F. — Vestnik Drevnei Istorii, No. 1. A Review. 1963.
4. Sanderson, I. The Dynasty of Abu. — N. Y., Knopf. 1962.
5. Feinberg, E. L. Obyknovennoe i neobychnoe (Ordinary and Extraordinary). — Novyi Mir, No. 8. 1965.
6. Berill, N. Worlds Apart. London. 1965.
7. Kondratov, A. M. Zvuki i znaki (Sounds and Symbols). — Znanie, 1966.
8. Frumkina, R. M. Statisticheskie metody izucheniya leksiki (Statistical Methods of Language Analysis). — "Nauka." 1964.
9. Norwood, V. G. Ch. Man Alone. — London, Boardman. 1958.
10. Shklovskii, I. S. Vselennaya, zhizn', razum (Life and Intelligence in the Universe). 2nd Edition. — "Nauka." 1965.
11. Porshnev, B. F. Vozmozhna li seichas nauchnaya revolyutsiya v primatalogii? (Are we Heading for a Scientific Revolution in Primatology?). — Voprosy Filosofii, No. 1. 1966.
12. Bel'kovich, O. M., S. E. Kleinenberg, and A. V. Yablokov. Zagadka okeana (The Mystery of the Ocean). — Molodaya Gvardiya. 1965.
13. Chauvin, R. Les sociétés animales de l'abeille au gorille. — Paris. Plon. 1963.
14. Hauser, P. (Editor). The Study of Urbanization, N. Y. 1965.
15. Rousseau, P. Histoire de la vitesse. — Paris, Press Universitaire de France. 1946.
16. Stine, G. H. Science Fiction is too Conservative. — Analog Science, Vol. 67. No. 3, N. Y. 1961.
17. Meyer, F. L'acceleration evolutive, Paris. 1946.
18. Bruner, J. Toward a Theory of Instruction. Belknap. 1966.
19. Zeman, J. Poznáni a informace. Praha. 1962.
20. Price, O. Little Science, Big Science. — N. Y., Columbia. 1963.
21. Dobrov, G. M. Nauka o nauke (The Science of Science). — Kiev, "Naukova Dumka," 1966.
22. Nalimov, V. V. and N. A. Chernova. Statisticheskie metody planirovaniya ekstremal'nykh eksperimentov (Statistical Methods of Optimum Experiment Planning). — "Nauka." 1965.
23. Lavrentiev, M. A. Berech' vremya uchenogo'. (How to Save the Scientist's Time). — Organizatsiya i effektinost' nauchnykh issledovanii, Novosibirsk, "Nauka." 1965.
24. Price, D. Regular Patterns in the Organization of Science. — Organon, No. 2, Warsaw. 1965.
25. Stroik, D. Ya. Kratkii kurs istorii matematiki (A Short Course in the History of Mathematics). — "Nauka." 1964.
26. Spearman, C. Psychology Down the Ages, Vol. 2. — London, Macmillan. 1937.
27. Walden, P. Chronologische Übersichttabellen. — Berlin, Springer. 1952.

28. Sirks, M.J. and Z.Conway. The Evolution of Biology. N.Y. 1964.
29. Haus, A. Sprachwissenschaft der Gang ihrer Entwicklung von der Antike bis zur Gegenwart. Freiburg. 1955.
30. Lyapunov, A.A.– Conference on Philosophical Aspects of Cybernetics, Moskva. 1962.
31. Ashby, W.R. Design for a Brain. 2nd Edition. – New York. Wiley. 1960.
32. Bir, St. Kibernetika i upravlenie proizvodstvom (Cybernetics and Industrial Control). 2nd Edition. – "Nauka." 1965.
33. Krushinskii, L.V. Ekstrapolyatsionnye refleksy kak elementarnaya osnova rassudochnoi deyatel'nosti u zhivotnykh (Extrapolation Reflexes as an Elementary Principle of Decision Making in Animals). – DAN SSSR, 121 (4): 762 – 765. 1958.
34. Rubinshtein, S.L. O myshlenii i putyakh ego issledovaniya (Intelligence and Ways of Its Study). – Izd. AN SSSR. 1958.
35. Lem, St. Dve evolyutsii – skhodstva i razlichiya (Two Evolutions – Similar and Dissimilar). – Nauka i Tekhnika, No. 8, Riga. 1965.
36. Clarke, A.C. Profiles of the Future. – Harper and Row. 1962.
37. Pushkin, V.N. Operativnoe myshlenie v bol'shikh sistemakh (Functional Intelligence in Large Systems). – "Energiya." 1965.
38. Pushkin, V.N. Evristika i kibernetika (Heuristics and Cybernetics). – "Znanie." 1965.
39. Wells, H.G. Anticipations of the Reaction of Mechanical and Scientific Progress upon Human Life and Thought. – New York and London. Harper. 1902.
40. Ryurikov, Yu. Cherez 100 i 1000 let (After 100 and 1000 Years). – "Iskusstvo." 1961.
41. Perel'man, R.A. Tseli i puti osvoeniya kosmosa (Means and Aims of Space Exploration). – "Nauka." 1967.
42. Zigel, F.Yu. Zhizn' vo Vselennoi (Life in the Universe). – Minsk, "Nauka i Tekhnika." 1966.
43. Biblioteka sovremennoi fantastiki (Library of Modern Science Fiction). Vol. 7. – "Molodaya Gvardiya." 1966.
44. Flammarion, C. Astronomie Populaire. – Paris. C. Marpon et E. Flammarion. 1881.

Chapter VI

SOME GENERAL TOPICS OF THE PROBLEM OF EXTRATERRESTRIAL CIVILIZATIONS

§1. INTRODUCTION

A new, scientifically minded approach to the problem of the existence and development of intelligent beings in the Universe is a definite possibility at this stage. There is no need to prove the scientific and the conceptual importance of further studies in this direction. The question of the possible existence of extraterrestrial civilizations has cropped up in one form or another throughout the history of science. This is a very difficult problem which embraces numerous fields and branches of sciences, so that at every particular level of scientific development, we can tackle only some of the aspects of the problem providing partial solutions.

The objective materialistic justification for the idea of multiplicity of inhabited worlds in its original form was the natural desire of man to penetrate the secrets of the evolution of human beings and human society, to reject the theological theses regarding the uniqueness of human life and intelligence in nature and the intrinsic difference between "soul" and matter.

Later hypotheses regarding the existence of other intelligent worlds were advanced in connection with certain scientific and technical advances. The wider boundaries of the visible Universe, pushed back by the rapid development of optical astronomy, provided ample food for thought on the subject of "ecological niches" which could sustain Earth-type life on other cosmic objects.

From the point of view of modern knowledge, these early hypotheses probably appear quite naive and unfounded, but the healthy methodological idea on which these hypotheses were based eventually led to the development of astrobiology and a new scientific discipline — exobiology, whose task it is to investigate the possible existence of conditions favoring the evolution of protein life forms in the Universe.

On the other hand, the problem of extraterrestrial civilizations often has been considered in connection with forecasts of the future development of human society. This topic excited the imagination of numerous scientists, philosophers, sociologists, and writers. K. E. Tsiolkovskii, in particular, was the first to point to the possibility of an "energy crisis" with further growth of industry, science, and technology, and he called attention to the inevitability of exploration and expansion into outer space.

These ideas, as we know, are adopted as starting premises in the great majority of modern work dealing with the problem of extraterrestrial civilizations.

At first glance, the natural and possibly the only approach to the problem of extraterrestrial civilizations would seem to stem from the two basic methodological ideas described above, which we shall call the "exobiological" and the "predictive" approach. Our reasoning will be based on the fundamental information available about the protein life form (the only one observed so far) and on the regular trends in the development of human society (again, the only known form of civilization at this stage).

There is, however, a possibility of a more general approach to the problem of extraterrestrial civilization. In this case, the problem of their existence is treated as part of a more general and complex problem which includes the study of the universal principles of structure, functioning, and evolution of complex large systems, with biological evolution and human civilization regarded as particular facets of such systems. This approach carries fundamentally new methodological features and is directly related to the development of the general theory of complex systems, which has recently received a considerable impetus from the direction of theoretical and technical cybernetics. This approach does not rule out the application of "exobiological" and "predictive" tools; it actually defines with greater precision their potential contribution to the problem of extraterrestrial civilizations.

The idea of a systematic approach to the problem of existence of extraterrestrial civilizations was clearly formulated by S. Lem /1/, whose arguments are often quoted in the following.

The problem can be formulated in a slightly different form: should we not try to analyze even now the fundamental principles of the problem of extraterrestrial civilizations in order to construct some "general theory of civilizations" based on the results of modern science? In our opinion, this is a very real possibility. It will enable us to clearly define the range of subjects that belong to the problem of extraterrestrial civilizations and thus define its exact position within the "general theory of civilizations" and also the position of the "general theory" within the general framework of classification of modern scientific disciplines.

The significance of the systematic approach clearly emerges from the examples of the fundamental difficulties which are encountered in the particular "astronomical" and "radio astronomical" aspects of the problem of extraterrestrial civilizations.

§2. THE METHODOLOGY OF THE "RADIO ASTRONOMICAL" ASPECT OF THE PROBLEM. THE "ENERGY" HYPOTHESIS

The advances in radio astronomy place us in a position where we can reasonably discuss the problems of detection of artificial signals from space in the radio spectrum. At this stage of our treatment, there is no need to formulate precise definitions of "signals" and "artificial origin." The exact meaning of these terms will be clear in each particular case from what follows, and certain improvements in the

definition will be introduced in § 3 of this chapter. The various ideas on which the analyses of the "radio astronomical" part of the problem are based are largely similar to one another. They are reviewed in some detail in the recent books by Shklovskii /2/, the article by Kardashev /3/, and in the two comprehensive collections "Extraterrestrial Civilizations" /4/ and "Interstellar Communication" /5/.

The search for signals from extraterrestrial civilizations is based on the following fundamental assumptions in these publications:

1) Radio frequencies provide the optimum range for transmission of meaningful signals over large distances.

2) Civilizations continuously increase their power requirements in the course of their development.

3) At a certain stage of development, civilizations inevitably start transmitting information into outer space.

4) The signals received from outer space can be decoded.

Not all of these assumptions are equally valid. The first of the four is apparently indisputable. It fully corresponds to the present-day level of our scientific and technical knowledge. We cannot envisage at this stage more efficient and practicable means of communication over interstellar distances. All other assumptions are highly speculative /2/.

The second point, despite its brief formulation, is characterized by a complex logical structure. First, it presents a definite prediction of the future development of human civilization. It is implied that scientific and technical progress will steadily continue in the direction of growing power and energy requirements and conquest of ever larger regions in space. Second, this principle is extended to all civilizations, or at least to a wide class of "anthropomorphic" civilizations in the Universe.*

The predictability of the future trends of scientific and technical progress is widely discussed in the current scientific press /6, 7, 8/. The concensus of opinion is that a comprehensive, systematic approach should be developed to the various problems of scientific and technical progress, assisted by a special scientific apparatus. The analysis should not be confined to the socioeconomic level of the factors of progress: adequate attention should be given to the general trends of development, the inner trends of the evolution of science and technology. This approach gave rise to a new scientific discipline – the "science of science" (see, e.g., /8/). One of the aims of this discipline is to devise a general theory of complex systems.

So far, the best examples of "long-range forecasts" of the development of human civilization are provided by science fiction writers. An analysis of the methods and techniques employed in the best products of this genre yields valuable information on the "psychology of forecasting." The main feature of these methods is the application of linear extrapolation into the future of those factors which are currently being implemented or are potentially ripe for implementation in the near future. S. Lem calls this technique "orthoevolutionary forecast." Direct time tests show that even

* The last sentence requires some qualification. It will be seen from the following that this statement /3 may be interpreted to have a two-fold meaning. On the one hand, it may imply that all the extraterrestrial civilizations are anthropomorphic, and on the other hand, we may intentionally limit the discussion to anthropomorphic civilizations.

the forecasts for the relatively near future greatly differ from the actual reality. The dialectics of growth, as we know, includes the quantitative changes as one (by no means principal) stage in the evolutionary process. In certain periods of development of human society, the discovery of new horizons (e. g., new forms of energy, new materials) produced a leap-like qualitative change in the methods of production, largely altering the way of life of the current generation and the further development of the society as a whole. Another shortcoming of the "orthoevolutionary method" is that it does not predict any alternative courses of development. One aspect of the phenomenon being considered is treated as an absolute factor, which is placed in antagonistic opposition to all the other alternatives. Actual evolution, on the other hand, is molded by an incessant interaction of polarities, and this interaction is one of the basic principles of the dynamics of progress.

A weak side of the "energy" hypothesis of the evolution of human civilization is its pronounced "orthoevolutionary" character. "It is tacitly implied that the rate of growth of the technical progress observed on the Earth during the last 200 years is a dynamically continuous process which can be arrested only by violent destructive forces ("degeneracy" or "suicide" of a civilization)" (/1/, p. 85).

The basic premises of the "energy" hypothesis are clearly based on known facts in the history of the development of science and technology during the recent period. The main motivation of this idea is the healthy desire to foresee and avoid the dangers of the forthcoming "energy" or "demographic" crisis, predicted in various sources /6, 7/. The conception of continuous expansion throughout outer space follows directly from the basic premises of the "energy" hypothesis and does not constitute a new additional assumption. However, the only cure that the "energy" hypothesis prescribes for these crises is a further quantitative step-up of power output. The search for new power sources and "lebensraum" is thus elevated to the pedestal of eternal problems. It is postulated that the current characteristics of the dynamic growth of humanity will persist for an indefinitely long period of time (in fact, hypotheses of this class maintain that this was the course of civilization from its very inception).* This principle applied to the energy hypothesis leads to the conclusion that the search for new power sources and free space places humanity in the uncomfortable position of striving to balance itself on a "razor's edge," since the slightest delay in making new power resources available will lead to catastrophic results. This explains the great importance attached to "space engineering" projects (Dyson's sphere, for instance) which constitute models capable of resolving the power and demographic crises. However, unlimited

* Note that, following in the steps of the originator of the idea of man's expansion into outer space, K. E. Tsiolkovskii, the Soviet authors generally associate this trend with their optimistic ethical-philosophical confidence in the unlimited potential of the human mind. On the other hand, Western scientists reveal a tendency to interpret the exodus into outer space as a result of the hardships of life and the conflicts of modern society. R. Simon /9/, for instance, tries to promote outer space, in the best Madison Avenue style, as a marvellous place for the development of private enterprise which is severely hampered on Earth. The main advantage of outer space, according to Simon, is its enormous "three-dimensional capacity." Therefore the "excess numbers" of humanity will spread to other planets, in order to escape from the congestion of the Earth, and thus enhance the "harmony" of human society.

"diffusion through space" and increase of power output increases the probability of other crises, which can be envisaged already at the present stage. We mean here the "information" and "organization" crises, which are ignored within the framework of the "energy" hypothesis. The authors who criticized the "space engineering" approach specifically mentioned obstacles of this kind. On the other hand, a detailed analysis of the means for overcoming these effects in the course of evolution leads to hypotheses which are radically different from the "energy" hypothesis /1/.

To summarize the preceding arguments, we would like to stress that the "energy" hypothesis is one of several alternatives based on the analysis of certain tendencies in the current development of humanity.

Let us now consider the universal applicability of the "energy" approach to the growth of other extraterrestrial civilizations. The Earth civilization, the only actual example before us, grows "technologically."* In theory, however, we should not reject the possibility of a "nontechnological" growth of a complex animate system /1/. A typical example of such growth is biological evolution, which takes the course of plastic adaptation to the environment. A system which develops in this way may reach a very high level of organization. From our point of view, however, it is not "intelligent." This conclusion is a fact for the biological evolution observed in this world. We can visualize, however, a directional activity taking the form of programmed autoevolution and introducing biological modifications intended to improve the adaptability to the environment. This civilization would appear very odd indeed from the point of view of the Earth civilization. This oddity, however, may be a direct result of our anthropocentric way of thinking, which automatically rejects the possible existence of other intelligent forms.

In any case, the definition of an "intelligent" system and to what extent it may be regarded as a "civilization" requires additional analysis.

The "energy" hypothesis prescribes one universal course of development for all the "technological," "anthopomorphic" civilizations.

Let us now consider the third basic assumption contained in the hypothesis regarding the feasibility of "radio astronomical" detection of signals from an extraterrestrial civilization. On the one hand, the third assumption is necessary to ensure a logical closure of the problem. If civilizations do not transmit radio signals, they cannot be detected by means of radio observations.** On the other hand, the third assumption nevertheless requires some logical justification.

We will naturally consider the case of a civilization intentionally transmitting information into outer space. This activity of extraterrestrial civilizations is generally justified on two counts:

a) it is assumed that the transmission of signals is related to the experimental, exploratory activity of an advanced "technological" civilization, trying to locate other similar civilizations /2/;

* Humanity took the course of active modification of nature, creating suitable conditions to sustain the parameters which are needed for its existence. This activity is far from "harmonic." Willingly or unwillingly, man destroys the natural ecological balance by this intervention, creating an artificial "ultra-low entropy" environment.

** This conclusion is fully valid in the "anthropocentric" statement of the problem. In general we should consider ways and means for detection of civilizations which do not transmit special signals announcing their existence but which are nevertheless "manifested" in specific forms of "behavior" (see §3).

b) it is assumed that the psychological-ethical trends of a highly organized civilization generate a certain pressure for the transmission of signals into outer space /3/.

Arguments of the first group clearly stem from the "energy" hypothesis applied to the growth of a "technological" civilization. "Technological" civilizations must explore the entire gamut of natural effects in the entire Universe around them. The most logical instrument for this exploration would be to establish a communication channel between neighboring civilizations. The "delay" in bilateral communication due to the tremendous distances in outer space does not constitute a fundamental difficulty in this treatment.

The second group of arguments is also sometimes quoted to justify the "voluntary" transmission of information to an unknown receiver. This factor is highly significant in calculations of the probability of detection of signals from supercivilizations /3/.

It can be argued that any functional activity of any complex system is justified only if it is essential for healthy growth and development of the system. In this sense, ethical and psychological factors are an outgrowth of deeper "behavioral principles" of a civilization. Therefore, the usual approach to the humanism or, conversely, the "aggression" of a civilization constitutes an entirely new factor added to the long line of previous assumptions, based on the extrapolation of current notions and concepts.

Let us now consider the possibility of decoding of the received signal. The logic behind the earlier attempts leads to the conclusion that an "anthropomorphic," "technological" civilization should transmit information in the form a semantic language system encoded in a certain form. Numerous attempts to construct formal languages for transmission of anthropomorphic concepts are known. One of these is Freudental's LINCOS /10/. The decoding approach is based on the assumption that the general system of concepts and knowledge is the same for the communicating civilizations. This assumption, however, is not a logical outcome of the preceding treatment. Gladkii /11/ pointed to the theoretical possibility of the existence of systems of knowledge with radically different elementary concepts for different civilizations. In these cases, the decoding of messages will naturally encounter serious difficulties.

For the purposes of our treatment, we should emphasize that the "radio astronomical" hypotheses of search for extraterrestrial civilizations are again committed to the anthropomorphic approach in their signal decoding attempts.

The analysis of assumptions which constitute the basis of /3/ and a similar group of hypotheses shows that the object of our search are signals from extraterrestrial civilizations which are close to the Earth civilization in all their basic activities and manifestations (including the details of the forecast growth).

It is naturally very difficult to follow blindly the "pure" principles of the form advanced in /3/. After all, we are dealing with working hypotheses and even their authors themselves continuously search for new ideas and methods of search for signals from extraterrestrial civilizations. In this sense, the aim of our critique is not to "reject" the current assumptions, but rather to define clearly the methodology behind these hypotheses.

A transitional stage toward new constructive possibilities of investigation is provided by the comprehensive discussion of the artificiality criteria of signals from outer space. Shklovskii proposed the conception of a "cosmic wonder" /2/. By "cosmic wonder" he understands the manifestations of intelligent activity on a cosmic scale, as observed by astronomical methods. Within the framework of the radio astronomical search for signals from extraterrestrial civilizations, the problem of discovery of the "cosmic wonder" sounds similar to the discovery of "call signals," i. e., signals which carry explicit information pointing to their artificial origin (see Chapter III). There is, of course, the question of unambiguous interpretation of this effect. For example, the decoding of a certain semantic system of signals received in the radio spectrum from an astronomical object would clearly point to the existence of a transmitting civilization. The artificiality criterion in this case was provided by the very decoding of the information contained in the signal. This, however, is an extremely lucky and quite unlikely turn of events. What are the other alternatives? According to Shklovskii, "... we can often detect distant supercivilizations purely objectively, by observations, because the associated objects do not follow the laws governing the behavior of inanimate matter or display remarkable and probably unnatural characteristics" /2/.

Statements of the kind "do not follow the laws governing the behavior of inanimate matter" and "remarkable characteristics" are ambiguous and uncertain. The logic behind the development of the natural sciences and their application to the study of astronomical objects imposes certain rigid restrictions on the possible interpretation of the most "outlandish" phenomena in the Universe as manifestations of "intelligent" activity.

Lack of precise criteria which distinguish the product of activity of a civilization from natural cosmic objects is conducive to unscientific speculations regarding the artificial origin of certain unusual phenomena (a well-known example is the unfortunate notoriety of the Tunguska meteorite) and, on the other hand, imposes an unnecessary "restriction" on the discovery and astronomical investigation of fundamentally new "natural" effects which enrich our knowledge in the fields of physics and other sciences.

A systematic approach to the solution of the problem of artificiality criteria will be described in § 3. Meanwhile we will consider another topic which is still the subject of lively discussion in the scientific literature concerned with extraterrestrial civilizations.

An important experimental fact is the conspicuous absence of "cosmic wonders." At this stage, we will ignore the possibility that this is due to our unreliable "artificiality" criteria and adopt the "anthropomorphic" approach. In accordance with the "energy" hypothesis, all civilizations should pass through a "technological" phase in their development, exploring and conquering the surrounding space. Many authors have noted (see Chapter I) that the rate of this technological development should be very high. In practice this means that the manifestations of cosmic activity of civilizations will be noticeable over periods which are very brief compared to the cosmic time scale. The absence of cosmic wonders within the framework of our hypotheses can therefore be attributed a) to the extreme rarity of civilizations, b) to the fact that all civilizations are roughly in the same "early" stage of development, and c) to the relatively short lifetime of civilizations. S. Lem examined these three possibilities critically.

What is the evidence in favor of the extreme rarity of civilizations? Baumshtein /12/ tried to prove the uniqueness of life on Earth by applying probabilistic and combinatorial techniques to calculate the likelihood of various ancillary conditions necessary for bio- and anthropogenesis (a certain gravitational pull of the Moon, a "required" succession of climatic conditions, etc.). S. Lem /1/ justly criticized the validity of the application of combinatorial methods to the highly complex dynamic evolutionary system. The evolutionary system is fundamentally plastic, so that the presence or absence of certain secondary conditions does not present it with the binary choice between "life and death" and only imposes certain restrictive trends on its future development. These calculations only prove the extreme unlikelihood of the development of an exact replica of the terrestrial anthropogenesis on other planets and thus place all the "anthropomorphic" hypotheses on shaky ground.

The words "extreme rarity" (if the other alternatives are ruled out) indicate an extreme dispersion of civilizations in the Universe, so that the Earth civilization is the only one occupying almost the entire visible Universe (otherwise, we would have witnessed the activity of super-civilizations of Kardashev's type). This assumption should be made consistent with the accepted cosmogonic concepts regarding the representativeness of the conditions in the solar system, the representativeness of the solar system in the Galaxy, and the representativeness of our Galaxy in the Metagalaxy.

The "rarity" of civilizations, however, is not sufficient to explain the total lack of observational evidence of their activity, unless we assume that the Earth civilization is unique.

Assumption (b) seems to contradict the cosmological information regarding the different age of the cosmic objects and the evolution of the Galaxy and the Metagalaxy. We can hardly accept the suggestion that the conditions favoring the evolution of life arose relatively recently on the cosmic time scale throughout the observable part of the Universe.

To justify the assumption of the brief lifetime of civilizations, various authors generally speculate about the possibility of catastrophes and crises which emerge in the course of accelerated growth /13/. However, the logic of these ideas in application to the conspicuous absence of "cosmic wonders" leads to the inevitable conclusion that all civilizations unavoidably perish in the early stages of their "technological" development /1/. It is only by adopting this fatalistic point of view that we can understand the lack of any signs of activity of "surviving" and rapidly advancing civilizations. This monstrous determinism is very difficult to accept without questioning.*

* We should emphasize that the idea of all civilizations perishing at a "convenient" moment, i.e., on the threshold of emerging into outer space or right before this phase, is particularly unlikely. In principle, the "death" of an individual civilization is far from contradicting the basic premises of the dialectic philosophy, which postulates that only matter as a whole is "indestructible," while all other phenomena originate, grow, and perish, giving room to new forms of life and existence /12/. A contrary point of view would have led to the erroneous statements that "the Universe is permeated with intelligence" or "intelligence is an indestructible attribute of matter." If we were to adopt these ideas, the lack of "cosmic wonders" would again lead us to the conclusion that the Earth civilization is unique in the entire observable Universe.

It is therefore more logical to assume that highly organized forms of existence regularly develop in different corners of the Universe. The spark of new life burns brightly, only to become extinguished and then reborn again under appropriate conditions. The span of life of these systems is difficult to predict. Modern science, in our opinion, does not provide even a rough estimate of the duration of a typical "psychozoic" era.

Proceeding from the absence of apparent signs of activity of extraterrestrial civilizations and the inherent weakness of the "anthropomorphic" hypotheses, S. Lem advanced an interesting hypothesis which maintains that the "nontechnological" evolution is characteristic for most existing extraterrestrial civilizations. According to S. Lem, the current "energetic" phase, including the expansion into outer space, constitutes only a very brief period in the life of a civilization, and it will eventually be replaced (in particular, under pressure from "information," "organization," and other crises) by a qualitatively new form of evolution. Lem's hypothesis naturally accounts for the absence of "cosmic wonders" and does not seek a definite answer to the question of the lifetime of civilizations. Note that from the traditional point of view we can hardly accept the idea of a "restriction" imposed on the expansion of a developing civilization into outer space. In any case, Lem's hypothesis contains fewer internal inconsistencies than any "anthropomorphic" hypothesis.

Let us now try to generalize our analysis of the weak sides of the theories of existence of extraterrestrial civilizations and the problems of communication based on "exobiological" and "predictive" principles.

A significant methodological shortcoming of the hypotheses of this group stems from an excessive abundance of additional assumptions. These a priori assumptions are associated with the "orthoevolutionary" reasoning of the authors (forecasting of future development based on linear extrapolation and "anthropomorphism," which maintains that this mode of growth is applicable to all (or most) extraterrestrial civilizations. According to this approach, certain facets of the phenomenon are invested with absolute importance and decisive significance, whereas other possibilities are ignored (e.g., the consequences of an "information" crisis).

As a result, theories developed in this way cannot resolve even the difficulties associated with those effects which are taken into consideration. Thus, for instance, the problem of the "energy" crisis is removed to outer space and its solution is postponed until "better times."

The "anthropomorphism" of our conceptions prevented a satisfactory development of the highly important notion of the "cosmic wonder." The artificiality criteria can be derived only from an internal system of anthropomorphic concepts. Therefore, the only clearcut criterion within this framework is a literal, word-by-word replication of external manifestations of human activity on other cosmic objects (e.g., adoption of an "anthropomorphic" semantic system of signals).

At the same time, the basis for the extrapolation of anthropomorphic hypotheses is provided by our knowledge of the protein form of life and the structure of the Earth civilization at the present phase of its development. Is there not an alternative course leading to a less controversial and more conclusive "theory" of extraterrestrial civilizations?

Any nonspeculative hypothesis should clearly rest on a foundation of scientific data. In this sense, we can maintain that the solution to the problem of extraterrestrial civilizations should be sought (at least at this stage) on Earth! Will it be enough to allow for all the present-day scientific concepts in the construction of this theory? Shklovskii /2/ pointed to the fundamental importance in the theory of extraterrestrial civilizations of such half-baked topics as functional definition of life and "intelligence." Shklovskii's argument sounds as if the respective studies are still in

their embryonic stage and have not yielded any tangible results which can be applied to the problem of existence and forms of intelligence in the Universe.

The theory of extraterrestrial civilizations is naturally in great need of exact definitions of life and intelligence, which are not restricted by any narrow particular model and the distinctive features of bio- and anthropogenesis. It is moreover clear that we are still very far from the development of sufficiently clear definitions of this kind. On the other hand, modern scientific disciplines related to cybernetic techniques indicate a new approach to the investigation of the surrounding reality. The fundamental nature of this novelty opens wide horizons in front of the corresponding branches of human knowledge. The new method of "cybernetic intelligence" is highly effective in problems dealing with the study of complex large systems.

The "cybernetic methods" enable us to introduce some order in the problem of extraterrestrial civilizations, to refine the terminology, to estimate the objectivity of the various statements, and finally to come up with a correct formulation of the basic problems. In what follows, we will try to present a systematic description of some constructive principles guiding the application of this method to the problem of extraterrestrial civilizations. We will also try to revise our attitude toward the various assumptions of the earlier theories. First, however, we will consider those publications in which the systematic approach has in fact been applied.

§3. AN ALTERNATIVE POINT OF VIEW. S. LEM AND HIS SUMMA TECHNOLOGIAE

In his book, Summa technologiae /1/, S. Lem treated in detail a number of topics associated with the problem of extraterrestrial civilizations. The "astronomical" aspect of the problem received only minor attention in this book. We have tried to show, however, that the problem of the existence of extraterrestrial civilizations is in fact part of a much wider problem concerned with the properties and the evolution of highly complex systems.

S. Lem deals with these "adjoining" problems and concentrates mainly on the possibilities of "forecasting" the future growth of civilizations. The principal features of the biological evolution are examined in detail in the light of modern scientific data. The potential possibilities of natural biogenesis and the control of "live" systems are discussed. These possibilities are compared with the requirements presented by science in connection with the design of complex artificial systems.

In /1/ it is noted that all the theoretical constructions dealing with the "forecasting" of the future development of mankind somewhat idealize the "thoroughness" of biogenesis. This is so because at the present stage mankind is still incapable of reaching the same degree of perfection in organic synthesis as the natural biological evolution has reached. We therefore tend to attach absolute importance to the gifts of nature, ignoring any possibilities of an "improvement" of human nature by artificial "autoevolution." "When chemical synthesis, the theory of information, and the general theory of systems reach a highly advanced stage, the human

body will appear in the light of these achievements as the most imperfect element." The next step naturally will be to seek ways and means for improving the "least perfect" element in the system of civilization. This will probably be achieved by "autoevolution," and not by improvement of the "conditions of life," since the very imperfection of the natural organization of the human organism sets a fixed limit to human life at around 100 years. It is frequently suggested that the human life span can be easily stretched to 150 – 200 years by appropriate medical treatment, but we doubt that this is indeed so. The basis for this suggestion stems primarily from the belief in the "enormous potential possibilities" presumably hidden in the human organism. In the process of evolution, the organisms reach a high degree of plasticity, adaptability, and build up a "reserve of reliability." The biological evolution, however, does not "plan" or "provide" for the future. Only the fittest survive, i.e., those organisms which are best adapted to the existing conditions.

Only those features of the biological evolution are selected which are significant for the ultimate "purpose" of genesis, i.e., for the survival of the species as a whole. As regards longevity, nature is "not interested" in the fate of the individual or how long he lives after having fulfilled the life functions significant for the continuation of the system (procreation, guardianship of the young generation). Accidental factors may combine to enable one individual to exceed a certain "necessary limit" of life expectancy. "Anthropomorphism" in the approach to the biological longevity of the human organism prescribes imaginary potential reserves to the bio-evolution, which presumably lie hidden until they are needed. In the final analysis this leads to teleological views on biogenesis. Moreover, the "methods" of evolution are purely statistical. The fact that certain individuals live beyond an average maximum age does not prove the longevity of the species as a whole. The fact that some people live to the age of 150 – 170 years is analogous to the fact that some people are geniuses: any suggestion that the entire population can be "educated to the level of genius" by an improvement of the conditions of life cannot be taken seriously.

An important place in S. Lem's book is devoted to the analysis of the "information" crisis. The danger of this crisis is considered in conjunction with the current trends in the development of science. The accelerated growth of the production resources (in particular, the search for new power sources) requires an ever increasing quantity of scientific information. This trend emerges from an historical example, analyzing the amount of research which had to be completed to ensure a transition from one kind of power to another. The corresponding amount of research steadily increased. The development of modern society will stop if the rate of data acquisition will cease accelerating. On the other hand, the "necessary" scientific discoveries cannot be planned or programmed. The "strategy" of science is essentially a matter of chance. In the course of scientific progress, the Earth civilization allocates scientific efforts to all the possible fields of research, since we do not know beforehand what turn the fundamental discoveries will take. This state of things naturally leads to an avalanche of new information, and also ties up a progressively larger number of people in scientific research. The high rate of growth should inevitably lead (in the light of quantitative treatment of the problem) to catastrophic results (e.g., depletion of human reserves that can be tapped by the needs of scientific research). The random character of the

"generation" of important discoveries prevents us from imposing any reasonable restriction on the scope of research. Given the present state of things, this would lead to even more serious consequences. Lem notes that even today excessive hyperbolization of individual fields of research, associated with rocket engineering and space exploration (the socio-political reasons are not to be ignored here), has a detrimental effect on basic research in other fields. And yet, different branches of science are not isolated or independent. Therefore, artificial retardation of the growth of some branches will eventually produce serious interference with further progress in the privileged branches of science also. Hypertrophy of the "popular" sciences often leads the scientist to lose much of the finesse of research and to try to solve all knotty problems by "frontal attack," by sheer quantitative step-up of the power level of the experiment (ever more powerful particle accelerators, ever larger radio telescopes, etc.) /14/.

How are we to avoid the "information" crisis? The development of science in human society reveals the particular importance of the ever increasing "information channel" between nature and the civilization. So far, the "channel" has been broadened by adding new branches of science and ever larger numbers of research workers. The advances in cybernetics give grounds for hope that in the near future we will be able to design complex machines to help man in the acquisition and processing of the increasing quantity of information. This does not indicate a change in the general character of scientific research, but only "automation" of the process.

It is clear, however, that by relying on "synthetic scientists" we only postpone the imminent crisis, without actually liquidating the factors responsible for the entire development. The only option (all other conditions being constant), according to S. Lem, is to create an artificial system which would directly extract the relevant information from the environment, i. e., a machine which would act not as a mere assistant for information processing, but as a powerful analytical system with capabilities far beyond those of the human brain. Lem develops the conception of a certain synthetic evolutionary system capable of increasing the "output" information in the process of its development, i. e., an arrangement not unlike the accumulation of genetic information in biogenesis, but with "directional and improved" action. A system of this kind will "cultivate" scientific results and conclusions. Lem shows that the feasibility of such a system does not contradict the premises of modern science and he proceeds to analyze the design details of the system from the point of view of the mechanism of scientific cognizance and the means of information transmission in biogenesis.

However, the creation of an autonomous data processing system will solve only part of the problems. The "information" crisis is not an independent phenomenon: it is conditioned on a whole range of other important satellite processes.

Lem also discusses the likelihood of other crises and catastrophic developments. Even if the problem of information acquisition finds a satisfactory solution, we will probably be far from a "quiet" mode of development in the form of "colonization of outer space" with gradual

expansion of the "living" space, while the population and the power resources will keep increasing continuously. One of the basic principles of existence and activity of complex systems is their controllability. We are not aware at this stage of the existence of a definite limit of structural complexity of a system (number of component elements). There is a possibility, however, that when the number of component elements exceeds a certain critical value, the system will become uncontrollable and disintegrate. In application to the "overgrown" civilization of the future, the problem of control no longer reduces to the banal and half-jocular question "how do we keep all the members of society busy?" We are dealing with such fundamental aspects, as, say, preserving the cultural unity of the giant system. An important cementing link in the development of humanity is the very continuity of the various stages, the transmission of "information" from generation to generation through vigorous exchange between the individual members of society. This process naturally enlarges the horizons of every individual. The impediment of information exchange in a giant supercivilization may lead to a loss of individuality, and every member will face the danger of becoming a highly specialized "cell" fulfilling narrow and limited service functions (no other "rational controllability" of the giant system can be visualized).

Lem notes that the concept of civilization is far from being synonymous with the free growth of all the possible individual freedoms. The opposite is probably true: the development of society imposes ever new restrictions, which are a necessary evil. This leads us to an interesting question: supposing scientific analysis confirms the unfeasibility of controllable systems made up of an excessive number of elements or shows that such a giant system can be rendered controllable only by bringing all the members of society to one common level, will this not be an excessive price to pay for the "freedom" of unlimited expansion into outer space?

Theoretically, we can reasonably assume that a society which has encountered fundamental difficulties on the way to expansion into outer space will reject this course of development. This need not indicate any "degradation" of the civilization. It has never been proved that the "spontaneous" growth of human civilization is the "best natural course" and will never lead to negative results or to tremendous irrational expenditures in the future. It suffices to mention the undesirable effects of the recent uses of atomic energy, such as the danger of genetic degeneration, or the harmful consequences of the wild, "unhusbanded" dissipation of the natural resources of the Earth.

Once recognized, the need is readily accepted by the civilization and is never regarded as an "unnatural" or "contra-natural" factor. For example, the "demographic" crisis is on the whole easily solved within the framework of ethics by birth control and family planning (a technique which is becoming progressively more popular with the growth of materialistic culture).

As an alternative to the "orthoevolutionary," "energetic" forecast of development, Lem advances a different hypothesis of his own. We have already mentioned the concept of "autoevolution." Another possibility, which enhances the "autoevolutionary" trend in Lem's opinion, is the creation of a "world within a world," i. e., a conglomerate of artificial conditions in a sufficiently large volume of space which is governed by a system of programmed artificial conditions characteristic of that "world."

The laws of motion, signal propagation, and structural elements of all the material objects in this "reserve" should be chosen so that they ensure optimum "control" of all the objects in that world, of the "imprisoned" civilization. This system would comprise an artificial machine consisting of two basic parts: the "environment" and the "civilization." The "function" of the machine amounts to the interaction between the two component parts, Lem in his Summa technologiae analyzes the feasibility of such systems.

Lem's hypothesis is of course highly speculative. However, it may lead to important methodological conclusions: besides the "energy" approach, there are alternative courses that a civilization may take, which theoretically are no less probable. The underlying idea of this treatment is that if the operating principles of complex systems can be disclosed, sooner or later the scientific progress will enable us to identify the optimum modes of development. Lem thus reduces the problem to its elemental level: what should be the "aim" of a civilization and what course of development should the civilization take in order to achieve that "aim"?

The "aim" of a complex system can be interpreted as the internally recognized principle of its action. We can speak of the objective categories of "intelligence," "conscience," "logic of systems," emphasizing that they are all functional properties of complex systems. The emergence of a certain "metaphysics" or "dogmatism" is inseparably linked with the practical activity of a complex system identified as a civilization. At every stage, the system does not have "complete knowledge" of the environmental reality, but it nevertheless functions as if its knowledge were complete. The functional determinism is associated with the conviction of its "correctness," as otherwise the system simply would not function. This is the basis of the natural "metaphysics" or "dogmatism." In the course of its development, a civilization progresses through a long succession of "dogmatic" or "working" hypotheses, which govern its functions and constitute a certain approximation to the objective reality. The process degenerates into "pure dogma" if the constant experimental checking and cross-checking against reality is stopped. In this case, all the objective events are distorted in the conscious mind and are dogmatically classified under one of the "working hypotheses"; the influx of new, additional information virtually ceases (an excellent example is the religious dogmatism). The strength of "working dogmas" is in their very mutability: they are constantly adjusted to fit the current level of science. The "aims" of a civilization can be determined only if complete information on the fundamental properties of complex evolving systems is available. Unfortunately, no such information is available at this stage. The rapid development of science gives grounds for hoping that in the near future the theory will be in a position to advance definition "recommendations" regarding the course of development of the entire human civilization. Implementation of these recommendations will be the task of a united, harmonically developing and "self-regulated" society.

We are in no position to choose between the two basic alternatives — expansion into outer space and creation of an artificial autoevolutionary Lem's world. It is easier to analyze the deficiencies of the various alternatives than to propose specific means of their implementation.

To conclude our brief review of Lem's book, we would like to emphasize again the great importance of the methodological approach advanced by Lem for the solution of a wide circle of problems related to the search for extraterrestrial civilizations.

The methods discussed in the previous part of the chapter are applicable not only to the "general theory of civilizations," i. e., the analysis of the fundamental properties of complex systems. This technique is also fruitful in application to "particular" problems. One such problem is the possibility of "natural" formation of complex systems in the Universe. A characteristic example of the "cybernetic" approach to the problem of the origin of life in the Universe is provided by Taube's work /15/. Proceeding from Lyapunov's functional definition of life /16/, Taube considers all the "natural" processes and material objects in the Universe which could provide the raw material for the creation of a living organism. Various necessary conditions are taken into consideration, such as sufficient abundance of certain elements, the ability of various compounds to combine into structures and to fulfill certain service functions (transfer of high- and low-entropy energy, transmission and storage of information). Taube came to the conclusion that the only material carriers of life under natural conditions are molecules of hydrogenous compounds which are spontaneously synthesized in an inanimate environment. The use of compounds without hydrogen, oxygen, and carbon as the "buiding blocks" of life forms is ruled out for fundamental reasons by the author. Let us analyze in some detail the validity of Taube's conclusions. First, Taube arrived at a precise formulation of the problem from the point of view of the functional principles of systems. He then investigated a large class of phenomena which could fulfill the functions of a "living system." Furthermore, he considered (although partially) the exact conditions under which complex "living" systems may originate in nature.

This approach is free from the "anthropocentric" bias of the studies concerned with protein life forms. In any case, Taube tries to prove the "universality" of the protein life form, as one of the very few permissible alternatives under the typical conditions prevailing in the Universe.

Taube's conclusions regarding the possible existence of "life" in the Universe are thus more objective. Using this approach, we can establish the possible "external morphology" of systems qualifying for the adjective of "living" and thus obtain more precise criteria for differentiating between the "animate" and the "inanimate" in the Universe.

One of the topics considered in connection with the problem of extraterrestrial civilizations is the possible impact of an encounter with intelligent beings from other planets. As a rule, the answer to this problem is formulated within the framework of "anthropomorphic" concepts. Extraterrestrial civilizations are considered with regard to their "humaneness" (or, conversely, "aggressiveness") /4/. The difference between civilizations is treated from purely quantitative "orthoevolutionary" aspects. Note that Stapledon /17/ was the first to consider in detail the problem of encounter with "differently made" civilizations.

A direct consequence of the "anthropomorphic" approach is the idea of "interplanetary aid" to be extended by civilizations following such an encounter.

In the light of our previous analysis, the encounter with other civilizations may be regarded as a characteristic "competition" between different intelligences. More "rationally constructed" systems are characterized by a higher adaptability, and this fact may have a decisive influence on other civilizations. The situation is not unlike that discussed on p. 248 in connection with the concept of "autoevolution." Having "become aware" of its "nonrational" makeup, a civilization will certainly put this information to work in order to improve itself. Failure to take any action because of "unacceptability" of the alternative courses of development of human society would be tantamount to the acceptance of the theological thesis concerning the uniqueness of humanity and its "predestiny." An outcome of encounters with extraterrestrial civilizations would therefore be acquisition of "purely scientific information" regarding comparative characteristics of the principle of action of other systems. From the methodological point of view, any results suppressing the anthropocentric elements in our scientific thought will be most valuable. Speculations on the subject of possible "conflicts" in interstellar encounters we leave to science-fiction writers.

§4. THE PROBLEM OF EXTRATERRESTRIAL CIVILIZATIONS FROM THE POINT OF VIEW OF THE GENERAL THEORY OF SYSTEMS

New scientific disciplines falling under the category of cybernetic methods developed as a generalization of principles which are still being used by science in the study of the reality around us. The analysis of phenomena of highly complex structure necessitated a revision of the basic principles of construction of scientific methods and analytical techniques. Therefore, the cybernetic approach does not introduce a "new way of thinking" into science, but the more accurate definition of the fundamental concepts opens a new and more effective way to tackle the most entangled problems in natural sciences. The value of correct methods of scientific research, even if they are confined to the rigidly "traditional" classical methodology, is in no way reduced by the discovery of new generalized principles. This idea regarding the continuity of scientific methods is best illustrated by the following example. In the previous part of the review, we criticized a certain conception of the problem of extraterrestrial civilizations. Our critique, however, did not weigh some hypothetical "cybernetic conception" against the "classical" approach, although in the course of the discussion we did mention the need for a systematic approach to the analysis of the problem. We mainly questioned the underlying premises of the "energy" hypothesis. On the other hand, the other particular problems relating to the existence of civilizations are solved correctly, and the "classical" solutions generally coincide with those stemming from "cybernetic" principles.

Thus, in his analysis of the distinctive features of the radio waves from suspected "artificial" sources, Siforov /18/ concentrated on the statistical structure of radio signals. He writes that "in particular, if the received signals are narrow-band signals, it is advisable to determine the two-dimensional probability density distribution of the end of the vector

describing the amplitude and the phase of the incoming oscillations in a plane. The surface describing these two-dimensional distributions provides an indication regarding the use of feedback in signal generation. It seems to us that the study of the statistical structure of the incoming signals will prove useful in deciding whether these signals are "artificial" or are generated by natural processes not related to the activity of intelligent beings."

We would like to call the reader's attention to the analogy between this approach and the method of the "black box" (p. 259), which is one of the fundamental techniques of modern cybernetics. A consistent application of this method in passing from restricted problems (the properties of the radio signal generator as inferred from signal statistics) to more general topics (artificiality criteria) appears to be quite promising.*

Before we advance further with our analysis, we shall have to introduce a number of concepts relating to "new" scientific disciplines, such as cybernetics, information theory, and others.

The term "system" in cybernetics represents an interrelationship of various elements which are described by sets of significant variables. Discovery of systems corresponding to this definition is linked up with the analysis of interrelated phenomena in the Universe. The main emphasis is placed on the principle of interrelationship, and not on particular cases of systems represented by certain material constructions. Examples of systems fitting this definition are the atomic nucleus, the solid state of an object, language, a game of chess, a conversation between two friends, etc.

An important point is the possibility of classification of systems. This classification is generally built according to the degree of complexity of the system, Moreover, the classification can be based on other principles also, e.g., systems may be classified according to the nature of the binding forces, namely deterministic and stochastic systems.

The development of the concept of a system and its properties leads to the definition of a "machine." A "machine" is a system whose state changes so that the state variables are interrelated by a certain transformation law. In accordance with the specific character of the system, we distinguish between deterministic and stochastic machines with simple and complex laws of transformation of the current parameters. A machine can be interpreted as a "target-oriented" system, i.e., a system whose organization is triggered in a sense to fulfilling the tasks that it is entrusted with /20/. The word "task" is not to be understood as "the goal set up by another system" (in particular, "man"), but only as the functional principle of the system.** Machines according to this definition cover a wide range of phenomena, covering atoms of the individual elements to planetary systems, cells and tissues of the living organisms, living organisms themsleves, the "biosphere," and even the biological evolution as a whole.

* This methodological approach to the problem of extraterrestrial civilization apparently was first advanced by Golei /19/.

* How to avoid in cybernetic treatment assigning conscious target or mission orientation to systems is a very important problem, not only from the viewpoint of the construction of a correct "metalanguage" for the description of some properties of complex systems. but also for elucidating the objective significance of such phenomena as "consciousness," "psychology," "purpose of existence" of a complex system (see /21/).

An extraterrestrial civilization may be treated as a system or a "machine."

An important characteristic of machines is the character of the dynamic coupling between the different parts of the system, e.g., the presence of positive or negative feedback.

S. Lem /1/ considered the remarkable classification of machines proposed by de Latille. De Latille distinguishes between three principal groups of machines according to the mode of their operation. The properties of the representatives of each successive class includes the properties characteristic of the previous classes. The first group of de Latille's classification (deterministic systems) includes simple and complex tools (non-automatic devices) and systems without feedback coupling to the environment. The second class includes organized regulated and self-regulated systems with feedback. This wide group of systems covers mechanical automatic regulators with feedback, programmed machines and self-programming installations (including man and animals, regarded as individual representatives). The third group includes systems which may change their structure and their "functional principles" using appropriate input material. The best example of this group is provided by biological evolution. De Latille also suggests the existence of a fourth group of systems, which are additionally endowed with freedom and ability to choose appropriate components from the environment in order to "build themselves up." The scientific and technical activity of mankind as a whole is obviously a system belonging to this group. In Lem's opinion, de Latille's classification can be extended to cover still another group of systems, namely those which do not select the input material for "self-organization" from the "naturally existing" resources in the Universe and do not apply the physico-chemical technology to manufacture the required synthetic materials, but rather create "synthetic" conditions which are never generated by natural physical processes. We are thus entering the domain of artificial creation of new forms of existence of matter, which according to Lem will be one of the attributes of mankind.

Classifications like the one above are of the greatest importance. They define the position of the system being considered among all the other systems using certain fundamental features, and thus permit formulation of the problem of analysis of the system in a suitable perspective. A clear formulation of the problem is especially essential in our search for extraterrestrial civilizations.

The next important step in the general classification of cybernetic concepts is the generalization of the concept of system stability. The principle of homeostasis plays an important role in this respect /20, 21/. According to the homeostasis principle, a target-oriented system functions in such a way that the values of certain significant internal variables are maintained between certain limits, despite a variety of (regular or irregular) external stimuli.

A homeostat according to this definition is a machine with an adequate regulating mechanism which sustains all the "critical life parameters" at a certain level. All the phenomena in the animate world are essentially homeostatic. The concept of homeostasis was actually introduced following a generalization of the results of biological observations. A discovery of "homeostatic behavior" in a system is therefore of the greatest importance for elucidating the exact nature of the particular phenomenon. Extraterrestrial

civilizations apparently also can be regarded as highly complex stochastic systems of homeostatic nature.

The fundamental nature of the homeostasis principle is further stressed by the fact that this is one of the very few clearly formulated conceptions which specify the probable "target" or "goal" of the evolution of complex self-organizing systems. Such properties of living systems as adaptability, survival, consciousness are on the whole governed by the principle of homeostasis. These properties therefore can be considered as a subsystem of the effective regulator in charge of sustaining the overall homeostasis of the system.

The concept of "intelligence" is of paramount importance, as we have seen, in any attempt to define effective artificiality criteria in the search for extraterrestrial civilizations. We have noted before that the methods of cybernetics provide a means for the construction of a functional definition of "intelligence." This definition emerges from the theory of complex self-programming and self-organizing systems, which are no longer very far beyond the reach of modern science.

A prerequisite of "intelligence" is primarily the ability of a system to store and process information. In the most general sense, information can be defined as a measure of ordering, a measure of the decrease in the uncertainty of the state of the system. In this sense, any machine has information, since its characteristic law of transformation limits the variety of other alternatives (states) which are thus unfeasible. Therefore any machine can be treated as an information processing machine /22/. This is indeed one of the fundamental principles of cybernetics. The modern theory of information deals with quantitative measurements of information and means of optimum information transmission. We are interested, however, in a slightly different aspect of the theory, namely what methods of information storage and processing are characteristic of "high-order" systems, i.e., what are the sufficient signs of "intelligence?"

One of the distinctive features of information transmission is that information is transmitted in coded form /20/. Information can be stored in the system in coded form, constituting a "memory bank" of the system. A suitable example is provided by the storage of genetic information in biological evolution. Proceeding from the available forms of information storage and transmission, we can move on to a more complex concept, that of the "logic of the system." The logic of a complex highly organized system is to be understood as its ability to reflect the external processes of the enivronment* by means of a certain set of internal responses presentable in a coded form and to apply these sets of states to the analysis and forecasting of external situations with the purpose of sustaining and "improving" the homeostasis of the entire system. The existence of a special "logic unit" is thus assumed, which operates with a set of coded symbols ("concepts"). The human brain is clearly one of these logic units. We are currently in a position to intelligently discuss the various "forms of logic" characteristic of complex automata, and we

* The very structure of the inrernal parameters may be interpreted by the logic apparatus of the system as a manifestation of an "external" situation, i.e., as an object for logical analysis (e.g., the study of human anatomy by man). This extension of the concept of "environment" is essential to avoid imposing restrictions on the possibilities of the analytical apparatus of the system.

are thus probably not far from a reliable classification of the distinctive features of the various "logics," according to the structure and the functional principle of the system.*

The "cybernetic" approach enables us to advance a definition of an "intelligent" system. We can tentatively define an extraterrestrial civilization as a highly complex stochastic machine of homeostatic character equipped with the required mechanisms in the form of "logic units" for information storage and processing, ability to analyze various situations and to apply the results of this analysis for purposes of directed evolution, in accordance with certain principles of directed action.

As we have stressed several times, the functional character of the cybernetic definition is the main feature. The feasibility (at least theoretically) of the functional definition forces us to advance a lucid formulation of our aims in the search for extraterrestrial civilizations.

The functional definition of a "civilization system" rules out the "anthropomorphic" approach.** The class of extraterrestrial civilizations encompasses not only "anthropomorphic" civilizations, but any other forms of "intelligent" existence, as long as they possess a sufficiently varied selection of parameters required for sustaining the programmed target-oriented activity. There is no more need for the various restrictions imposed on the search for the possible manifestations of "intelligent activity" by a certain class of typology, and the ambiguity of the statements regarding the degree of reliability and single-valuedness of the interpretation of critical "difficult-to-explain" phenomena is automatically eliminated. The same naturally applies to the artificiality criteria of radio signals. In principle, this presents us not only with an opportunity to "decode" semantic information, but also to determine the origin of the signal by gradually refining the methods of structural analysis of the signal.

* This line of reasoning shows that we will be hard pressed indeed to define a clearcut boundary between "intelligent" and "unintelligent" systems. This is further borne out by some findings of modern biology, which point to the existence of certain elements of "consciousness" and "logic" in various animal species. This only provides additional proof of the functional character of the very concept of "intelligence," which is based on a purely material foundation — the structure and the presence of certain mechanisms. Such concepts as "consciousness," "emotional response," etc., are related to certain properties of complex systems 21/.

The traditional intuitive approach maintained that "consciousness," "logic," and "emotions" are the principal and decisive attributes of an "intelligent" system. Modern science opens new ways for the interpretation of the strictly "utilitarian" significance of such aspects of civilizations as religion and art. Art in the light of the theory of complex systems may be interpreted not only as a means for acquiring additional information, but as a "teaching" or "training" process regulating and controlling the "emotional" and "aesthetic" properties of the complex highly organized systems /1/. In any case, "tuning" phenomena in the complex system corresponding to human civilization may and should be interpreted as phenomena which are objectively connected with the "functional principle" of the system, and their properties should be elucidated through a study of their functional significance for the system /1/. All of historical materialsim is based on this point of view.

** In any case, the "anthropomorphism" of the relevant statements is "lowered" to such a level that we can reason "nonanthropomorphically" concerning the laws which govern the world around us. Human consciousness typically analyzes the world by means of a certain logic apparatus (which includes, e.g., "mathematization" of the methods of analysis). To speak of "anthropomorphism" at this level is to maintain that the outside world is arranged "chaotically," without any "causal relations," etc. These statements clearly contradict the materialistic theory of knowledge.

The functional definition of "civilization" suggests some general principles for the treatment of the "theory of extraterrestrial civilizations." The problem is reduced from one of "astronomical" importance to a typical "terrestrial" problem. The advances in the theory of complex systems regarding the fundamental properties of highly organized forms of existence should provide a proper foundation for the development of a valid "particular" theory of highly organized extraterrestrial systems. In this respect, the importance of theoretical cybernetics in the study of extraterrestrial civilizations is analogous to the contribution of theoretical physics, say, to modern astronomy. The theory of extraterrestrial civilizations, on the other hand, may contribute to the development of cybernetic concepts, e.g., through analysis of the specific conditions prevailing on various cosmic objects.

The cybernetic techniques apply to a wide range of effects. In cybernetics, any complex system can be studied by the "black box" method.

A "black box" is a mode of a system which is to be studied without any information being available beforehand about its internal structure. This system can be simulated by a machine with an "input" and an "output." The "input" is the complete range of stimuli and interactions to which the system is exposed, whereas the "output" comprises the various responses of the system to specific input stimuli. In principle, a real system may have an infinity of "inputs" and "outputs." The number of "inputs" and "outputs" in a certain sense is determined by the method and the latitude of the experiments to which the object is subjected.

Before proceeding with a detailed discussion of the "black box" technique, we would like to discuss briefly the highly fruitful concept of models in science. F. Engels, more than a century ago, published a brilliant analysis of the genesis of scientific knowledge as related to the practical activity of mankind and called attention to the fact that the unknown or ungrasped phenomena, which are "things in themselves," are converted into a fully known "thing for us" once we succeed in reproducing the corresponding phenomenon artificially. The transformation from a "thing in itself" into a "thing for us" is a lengthy step-by-step process in the course of which we investigate one after another the various new features of the phenomenon, gradually approaching full knowledge of all the basic properties of an objectively existing "thing in itself." This reasoning is the basis of the modern scientific concept of a "model." The methods of modern science (e.g., physics) almost invariably make use of particular models of the phenomenon for purposes of mathematical description. All the theories of modern physics are essentially based on certain physical and mathematical models. If the model fully corresponds to the original and it covers all the properties of the source object, the model is said to be isomorphic to the original. In this sense, for example, we can say with complete certainty that the representatives of one class of objects are fully isomorphic to one another. Thus any two hydrogen atoms are fully isomorphic to each other. From the standpoint of cybernetics, however, a model is isomorphic to the original object as soon as it faithfully duplicates all the operations and functions that the original object performs; there is no need to demand complete similarity of the model and the object in cybernetics.

Models of complex systems generally are not fully isomorphic to the real object. This leads to the concept of homomorphic models. A homomorphic model corresponds to the original phenomenon to a degree, on a certain level, and provides a correct interpretation of a limited range of properties of the original phenomenon. For instance, an electronic computer is homomorphic on a certain level to the human brain, since it performs a number of definite logic operations, although functionally the computer is basically different from the human brain. It is significant that were we able to devise a machine capable of performing all the various operations of the human brain, so that the "potentialities" of the machine and the brain would be identical, the result would be a cybernetically isomorphic model of the brain. This again emphasizes the special importance attached in cybernetics to the functional principles of processes, rather than to the particular material expression in the form of a "thing" with all its various external signs and features.

We are now ready to present the principles of the "black box" technique. The "black box" approach is applicable to highly complex methods whose structure is inaccessible to direct study.

Applying certain stimuli to the system input (which is equivalent to the various interactions of the system with the environment and with other systems, which need not be artificial "experimental devices"), we can study their functional relationship to the output responses of the "black box." At every stage, a model homomorphic to the actual phenomenon is created (e.g., in the form of a working hypothesis). The principal aim is to establish the law of transformation from "input" to "output," i.e., the functional principle of the machine. Accumulation of data permits the construction of homomorphic models on progressively more sophisticated levels. In this limit, an isomorphic model of the phenomenon will be obtained. The basic features of this approach lead to an apparently paradoxical conclusion: in principle, full and exhaustive information can be obtained about the "black box" without in any way disclosing its actual physical structure. The missing link, however, emerges directly from the definition of an isomorphic model. An isomorphic model is functionally equivalent to the real system. This model is completely interchangeable with the real object, since it performs all the analogous functions. The construction of an isomorphic model corresponds to artificial duplication of the real phenomenon. It was Engels who first suggested basing the criterion of transformation from a "thing in itself" to a "thing for us" on the possibility of artificial duplication!

To apply the "black box" approach to the problem of extraterrestrial civilizations, we have to discuss the information flow process in observations using the "black box."

In addition to the above, we should remember that a flow of information is possible only in certain systems, where the different system components are linked by communication channels.

The "black box" technique can be represented in the form of a certain information machine, where the object and the observer constitute a single system with a feedback loop (Figure 72). It is significant that the information is transmitted in coded form. Therefore, if the different system components use different codes, suitable code-translating units should be provided, converting the code of one component into a code "understandable" to the

other component. Figure 72 thus shows a block diagram of an information processing cycle in the "black box" technique. The stimulus applied at the "black box" input determines the output response in the form of a certain signal.* The observer "decodes" these signals and interprets them accordingly. The feedback loop is provided by the observer who tests his hypotheses and conclusions by applying new stimuli to the system input. The choice of the input stimuli delivered to the "black box" is essentially a process of translation of the "signal" originating from the observer into a form "understandable" to the "black box." Each information exchange cycle provides some additional data to the observer who, having accumulated a sufficient quantity of information, will create a "black box" of an appropriately high level.

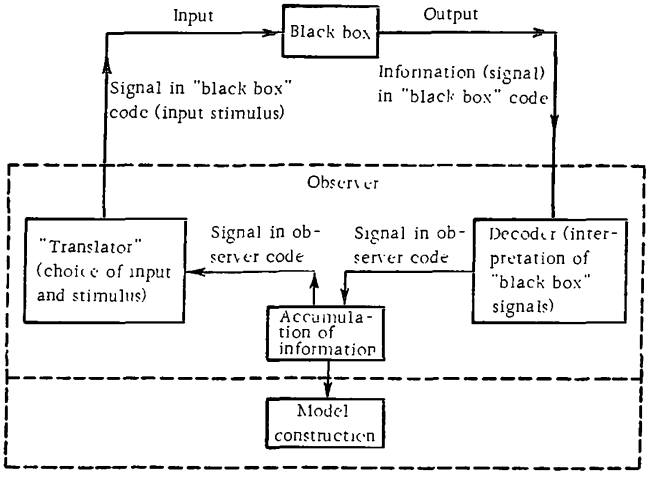


FIGURE 72. Observations by the method of a "black box" with an input, diagramed in the form of an information machine.

This constitutes a schematic description of a "learning model," illustrating the process of analysis of an unknown effect by the usual scientific methods.

The "black box" approach is particularly useful for the analysis of phenomena on distant extraterrestrial objects. The problems of astronomy essentially reduce to the investigation of inaccessible "black boxes." Astronomers of all ages seem to have been applying this technique, without realizing the cybernetic significance of what they were doing.

* The concept of a "signal" is a fairly complex one, and it often lends itself to an ambiguous interpretation. In the most general sense, "signal" is to be understood as a mode of information transmission. This definition does not restrict us to a material information carrier, a "material" interaction between the two parts of the information machine. For example, in an "intelligence machine," the lack of interaction is also a kind of signal, since it indicates (carries information about) the absence of certain particular phenomena. (For example, the absence of "cosmic wonders" is a significant piece of information in the theory of extraterrestrial civilizations.)

In astronomy, however, the "black box" problem is somewhat more complicated than before. The astronomical objects in a sense are "black boxes" without "input." Because of the tremendous distances to these objects in space, we only detect the output signals, which are invariably in the form of electromagnetic radiation. The astronomer cannot "experiment" with the object by altering the conditions of the phenomenon. Thus there is no feedback in the "observer–object" system (Figure 73).

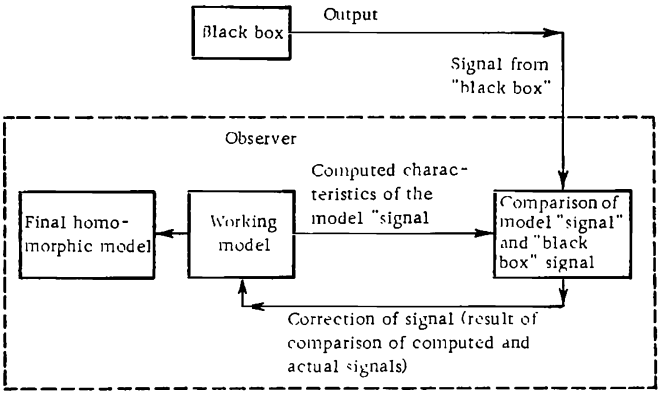


FIGURE 73. Observations by the method of a "black box" without an input, diagramed in the form of an information machine.

The cyclic flow of information in this machine is confined to the subsystem "observer." The output signals of the "black box" are compared with those which would presumably constitute the output of some hypothetical model of the phenomenon. In actual practice, the entire astronomical reasoning is based on analogies. The multivalued choice of models may be limited by analyzing classes of analogous phenomena. The "predictability" of new aspects and features of the phenomenon plays a very important role in the process of decision making. However, from the point of view of our cybernetic concepts, the astronomical research methods essentially amount to successive "rejection" of inappropriate homomorphic models. The construction of the isomorphic model, on the other hand, encounters fundamental difficulties. The lack of the "observer – object" feedback limits the procedure to the construction of a homomorphic model of a certain level. As a result, the astrophysicists are forced to consider the new possibilities of signal detection from extraterrestrial objects and to improve the existing methods. Particular stress has been placed recently as a result on the analysis of "fine" effects.

The neglect of the basic fact that no isomorphic models exist in astronomy often places the astrophysical theories on an excessively speculative basis. On the one hand, the astrophysicists often have to reluctantly abandon their traditional conceptions; on the other hand, they are much too keen on "fashionable" effects and attempt universal interpretation of the various phenomena from the standpoints of the particular theories which happen to be in vogue at the particular time.

Modern cybernetics still has not developed a theory of analysis of "black boxes" without an input. It is, however, advisable to introduce some of the methodological ideas of this science into astronomy. It may yield an additional criterion for testing the reliability of the multitude of theoretical assumptions regarding our knowledge of the Universe. Some recent publications show applications of the "quasi-cybernetic" approach to the analysis of certain astronomical phenomena.

Thus Gudzenko and Chertoprud /22, 23/ tried to investigate solar activity by a method based on the analysis of the statistical properties of the "signal," namely the time-dependent parameters of solar activity. This analysis elucidates the functional principles of mechanisms responsible for the appearance of solar activity (e.g., whether or not this is a self-sustained oscillatory system). This approach, in our opinion, is markedly superior to the traditional observational methods, which primarily search for particular carriers of activity and only then try to fit it with a plausible model explaining the functional features.

The cybernetic approach is of special significance in the problem of extraterrestrial civilizations. In principle, progressively more detailed studies of the structure of signals from a "no-input black box," based on the block-diagram of Figure 73, produce homomorphic models of progressively higher levels. Our understanding of the functional principles of the complex system is improved correspondingly. Theoretically, we can construct a classification typology of functional properties of progressively increasing complexity; this approach permits assigning every individual object to a certain class of the classification.

For example, a very extensive class of objects comprises systems with feedback of all kinds and of all degrees of structural complexity. The appropriate data can be obtained even now, by a detailed analysis of the statistical structure of the incoming radio waves, say. The next stage would be to try and define a narrower subclass of objects displaying homeostasis. Finally, a group with even more complex functional properties will be isolated from the homeostatic class. Ultimately, we can visualize in principle a class of objects which are homomorphic on a sufficiently high level to the Earth civilization. This is a fantastically difficult job, and it will not be solved in less than a few years or even decades. The attractive prospect of this approach, however, is that it permits formulating in precise and consistent terms the actual purpose of research. This approach eliminates the uncertainty inherent in the interpretation of effects within the framework of their classification into "artificial" and "natural." Each effect will now be regarded as representative of certain particular features of the generating mechanism. If we can prove that these "features" define a certain "logic system," we shall have discovered an extraterrestrial civilization.*

* This method is intrinsically "minimalistic" in its evaluations. Having established that a certain object can be classified in terms of the distinctive features of its output signal in the lowest class of the typology, we conclude that this object definitely belongs to that class, without, however, ruling out the possibility of its being part of some higher class of the same typology. This frame of reference is clearly the most adequate for astronomical research, because of the impossibility of proving the isomorphism of the result to the actual phenomenon.

Finally, the class of "civilizations" in principle may contain systems with greatly differing morphological features (e.g., theoretically we can envisage "non-technological" systems with an entirely different set of elementary concepts, etc.).

It is naturally important to take into consideration all the additional information concerning the properties of the proteinic life forms, the specific conditions in space in different parts of the Universe, and a variety of other data which are currently used to a varying degree in the formulation of the problem of extraterrestrial civilizations (including the search for "anthropomorphic" civilizations!).*

In principle, the search for extraterrestrial civilizations should proceed according to the following methodology.

We have to concentrate on the various aspects relating to the functional principles and the basic laws of behavior of very complex systems (covering, in addition to structural analysis, the problems of evolution and forecasting of the future forms of existence of the system). The aim is to create a reliable classification of systems according to significant distinctive features. All these are topics which fall within the competence of "terrestrial sciences," specifically the theory of complex systems. A very difficult problem is the determination of the "artificiality criterion," i.e., a system of signals which can be identified unambiguously as originating from a highly organized system ("civilization").** These signals should have certain distinctive features which can be applied to differentiate them from other signals, however complex, which originate from astronomical objects that cannot be regarded as "civilizations." From the point of view of cybernetics, this amounts to the determination of the level of organization of a "black box" without input from its output signal /24/ or the synthesis of the originating system from the characteristic features of the observed signal.

Certain signal sequences (structures) can be theoretically devised such that the originating system will of necessity have a number of highly complex features (e.g., a "memory," ability to "recognize patterns," create abstractions, etc.). These signal sequences constitute regular structures organized in a special manner according to definite set-theoretical principles, as if to "demonstrate" certain functional properties of the system enabling it to perform complex operations /25/. The development of these topics is again not a strictly astronomical problem.

* There is a possibility that the search for an "anthropomorphic" civilization will culminate in the discovery of a semantic system of communication signals. This possibility does not detract from the generality of our conclusions. The above reasoning should not be interpreted to indicate that civilizations of this kind cannot be discovered, since after all the subclass of "anthropomorphic" systems is part of the corresponding cybernetic typology. If a "man-like" intelligence were to be discovered in the form predicted by the "anthropomorphic-energetic" hypothesis, this would justify the assumption of the universal applicability of the "technological" and energetic mode of development, but only on grounds of correspondence to some deeper underlying principles of the growth of complex systems.

** Here we naturally ignore the question of the existence of "anthropomorphic communication" with semantically decodable information. This particular case provides an unquestionable "artificiality criterion." However, "semantic communication" requires sufficiently long "messages" adequate for successive decoding of the individual symbols, so that we again return to the problem of "call signals," i.e., sufficiently brief endings which are highly effective for detection purposes (see Chapter II).

Depending on the success of the above measures, we will have to continue with a consistent study of the "functional principles" of astronomical objects, after clearly formulating exactly what properties we are interested in discovering. This is evidently an astronomical problem. It also includes a generalizing part: elucidating the effect of various cosmic conditions on changes in functional principles of complex systems.

Other branches of science, e.g., radio astronomy, certainly can make their contribution to the analysis of individual problems. The study of the various "noises" interfering with proper signal detection and distorting the signals constitutes a separate part of the comprehensive overall investigation.

All the available scientific data lead to the conclusion that a precise methodology can be devised for the solution of the problem of extraterrestrial civilizations on the current level. The great complexity of the problem stems from the fact that it is inseparably linked with even more fundamental problems. Therefore, only further advances in the methods of cybernetic analysis, the general theory of systems, biology, and other disciplines will enable significant progress to be made toward the solution of the problem of extraterrestrial civilizations. We cannot rely on a "lucky chance" that will enable us to "guess" the answers to the main questions, which are not even always clearly formulated. The same also applies to experiments aimed at detection of astronomical signals bearing signs of "artificial" origin. Here the primary problem is clear and precise formulation of the "artificiality criteria" and detailed analysis of a very extensive class of astronomical phenomena. On the whole, this is not a fundamentally new problem. S. E. Khaikin notes that the problem of systematic search for radio signals of artificial origin on the whole coincides with the fundamental problem of radio astronomy: accumulation of information about the cosmic radio sources /26/. The differentiation will become possible only after the application of "artificiality criteria" to particular classes of objects.

The problem of the "general theory of civilizations" will clearly be one of the major subjects of contemporary and future science. Any progress toward the solution of this problem is predicated on the general advancement of science. There is no doubt that this field of research will eventually occupy a prominent position among the other scientific disciplines.

Bibliography

1. Lem, S. Summa technologiae. – Kraków, Wyd. Lit. 1964.
2. Shklovskii, I. S. Vselennaya, zhizn', razum (Life and Intelligence in the Universe), 2nd Edition. – "Nauka." 1965.
3. Kardashev, N. S. – Astron. Zhurnal, Vol. 41 : 282. 1964.
4. Vnezemnye tsivilizatsii (Extraterrestrial Civilizations). Proceedings of a Conference, Byurakan, 20–30 May 1964. – Izd. AN Arm. SSR. 1965.*
5. Cameron, A. (Editor). Interstellar Communication. – New York. Benjamin. 1963.
6. Lilley, S. – Sociology of Science, N.Y. 1962.

* [See footnote on p. 11.]

7. Price, D. — Discovery, Vol. 6 : 240. 1956.
8. Dobrov, G. M. Nauka o nauke (The Science of Science). — Kiev, "Naukova Dumka," 1966.
9. Simon, R. — Astronaut and Aeronaut, Vol. 3 : 59. 1965.
10. Freudental, H. Lincos, Amsterdam. 1960.
11. Gladkii, A.V. — In: /4/ : 145.
12. Baumshtein, A. I. — Priroda, No. 12. 1961.
13. Breisuell, R. In: /5/ : 271.
14. Hoyle, F. Of Men and Galaxies. Univ. of Washington Press. 1966.
15. Taube, M. Hydrogen the Carrier of Life. — Nucl. Energ. Inform. Center, Warsaw. 1965.
16. Lyapunov, A.A. — In: "Kibernetika, myshlenie, zhizn'," p. 127. "Mysl'." 1964.
17. Stapledon, O. Last and First Man. — London, Peng. B. 1939.
18. Siforov, V. I. — In: /4/ : 121.
19. Golei. — see /5/.
20. Ashby, W. R. An Introduction to Cybernetics. — Chapman and Hall. 1956.
21. Bir, St. Kibernetika i upravlenie proizvodstvom (Cybernetics and Industrial Control). — Fizmatgiz. 1963.
22. Gudzenko, L. I. and V. E. Chertoprud. — Astron. Zhurnal, Vol. 41 : 597. 1964.
23. Gudzenko, L. I. and V. E. Chertoprud. — Astron. Zhurnal, Vol. 43 : 113. 1966.
24. Panovkin, B. N. — Doklad na Sessii NTORiE im. A. S. Popova, Moskva, May 1967.
25. Panovkin, B. N. — Doklad na 1-oi Konferentsii po kosmicheskoi radiosvyazi. Moskva. 1968.
26. Khaikin, S. E. — In: /4/ : 83.

Printed in Jerusalem, Israel

TT 70-50081
NASA TT F-63